U. S. ARMY RESEARCH OFFICE

Report No. 76-1

February 1976

TRANSACTIONS OF THE TWENTY-FIRST CONFERENCE OF ARMY MATHEMATICIANS

Sponsored by the Army Mathematics Steering Committee

Host
U. S. Army White Sands Missile Range
White Sands Missile Range, New Mexico
14-16 May 1975

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, North Carolina

FOREWORD

These Transactions preserve in print most of the invited addresses and contributed papers of the Twenty-first Conference of Army Mathematicians. These meetings are sponsored by the Army Mathematics Steering Committee (AMSC) on behalf of the Office of the Chief of Research, Development and Acquisition. Members of this Committee request that the guest lecterers be effective researchers who are in frontier fields of live current interest. They feel that the addresses by the invited speakers as well as contributed papers by Army personnel will stimulate the interchanges of ideas among the scientists attending said meetings.

In June of 1973, Messrs. W. A. McCool and William Sheperd of the White Sands Missile Range called the Army Research Office (ARO) to inquire about the possibility of holding either the Twentieth or the Twenty-first Conference of Army Mathematicians at their installation. Since the next, the Twentieth Conference, had been scheduled for the US Army Natick Laboraotires, these gentlemen were thanked and advised that a written offer to host the 1975 meeting would be appreciated. Dr. Richard H. Duncan, Technical Director and Chief Scientist of the White Sands Missile Range, in a letter dated 10 May 1974 issued a formal invitation to hold this meeting at his installation. Part of his letter is quoted below:

> I cordially invite the Applied Mathematics Subcommittee, US Army Mathematics Steering Committee, to hold the Twenty-First Conference of Army Mathematicians at the US Army White Sands Missile Range, NM, in May 1975. We would be pleased to serve as host to the conference.

I believe a conference held here would attract a number of Western Army Mathematicians who do not normally attend the conferences. Their attendance and participation would be of substantial benefit to the Army installations in our area.

If the conference is held here, Mr. P. J. Higgins, Chief of the Analysis and Computation Division, National Range Operations Directorate, will be in charge of local arrangements. Please contact him for additional information.

Over 115 Army and academic scientists attended this 14-16 May 1975 conference, which convened in El Paso, Texas. It is pleasing to note that over hald of these individuals were from the host installation. On the afternoon of the second day of the meeting, Mr. P. J. Higgins, the Chairman on Local Arrangments, arranged a trip to the White Sands Missile Range and a very interesting tour of that base. Those in attendance are indebted to him and members of his staff for this three day meeting.

The success of this conference was due to many persons, including the active and enthusiastic members of the audience, the chairmen of the sessions, and authors of the many papers. The members of the AMSC would like to thank these gentlement for taking time to prepare papers for these Transactions so that many persons unable to attend this symposium can profit by their contributions to the scientific literature.

TABLE OF CONTENTS*

*This table of contents contains only the papers that are published in this technical manual. For a list of all papers presented at the Twenty-first Conference of Army Mathematicians, see the program of the agenda.

PROGRAM

21st CONFERENCE OF ARMY MATHEMATICIANS
US Army White Sands Missile Range, New Mexico

All general and technical sessions will be held in the Rodeway Inn (Bassett Center) 6201 Gateway West, El Paso, Texas.

Wednesday, 14 May 1975

0820-0845 REGISTRATION - ERICA ROOM

0845-0900 OPENING OF THE CONFERENCE WELCOMING REMARKS - ERICA ROOM

0900-1000 GENERAL SESSION I - ERICA ROOM

SPEAKER: Professor Alan J. Hoffman
Mathematical Sciences
IBM Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, New York 10598

TITLE: Graph Theory and Eigenvalues of Matrices

CHAIRMAN: Dr. Ivan R. Hershner, Jr.
Room 3E 453
The Pentagon
HQDA (DAMA-ARZ-D)
Washington, D. C. 20310

1000-1020 BREAK

1020-1200 TECHNICAL SESSION I - ERICA ROOM

CHAIRMAN: Dr. Edward W. Ross, Jr.
Staff Mathematician
US Army Natick Laboratories
Natick, Massachusetts 01760

1020-1200 TECHNICAL SESSION I (Continued)

EFFECTS OF CONCENTRATED MASS AND THRUST DIRECTIONAL CONTROL ON THE STABILITY OF FREE-FREE BEAMS
Julian J. Wu, Bénét Weapons Laboratory, Watervliet Arsenal, Watervliet, New York

THE STANDARD LINEAR MODEL IN THE STABILITY AND MASS OPTIMIZATION OF NONCONSERVATIVE EULER BEAMS
Charles R. Thomas, Bénét Weapons Laboratory, Watervliet Arsenal, Watervliet, New York

BUCKLING OF ORTHOTROPIC RECTANGULAR CYLINDERS
Earl C. Steeves, US Army Natick Laboratories, Natick, Massachusetts

ENERGY RELEASE RATE IN TERMS OF COMPLEX ANALYTIC FUNCTIONS
M. A. Hussain and S. L. Pu, Watervliet Arsenal, Watervliet, New York

1020-1200 TECHNICAL SESSION II - TROPHY ROOM

CHAIRMAN: Dr. Badrig M. Kurkjian
AMCRD-R
5001 Eisenhower Avenue
Alexandria, Virginia 22304

DEVELOPMENT AND APPLICATION OF DYNAMIC MATHEMATICAL MODELS FOR EVALUATION OF MILITARY SYSTEMS, FORCES AND DOCTRINE
Roger F. Willis, US Army Combined Arms Combat Development Activity, Ft. Leavenworth, Kansas

APPLIED TECHNIQUES OF ACQUISITION AND PRODUCTION COST ANALYSIS
Van A. Puryear, US Army Troop Support Command, St. Louis, Missouri

HOMEOSTATIC CRITERIA FOR ASSESSMENT OF THE MORPHOLOGICAL STATE OF LARGE SCALE HYBRID ANALOG-SIMULATION COMPUTER MODELS
Howard M. Bratt, US Army Air Mobility Research and Development Laboratory, Ft. Eustis, Virginia

Wednesday PM

1200-1300 LUNCH

1330-1500 TECHNICAL SESSION III - ERICA ROOM

CHAIRMAN: Professor Ben Noble
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, Wisconsin 53706

THE APPLICATION OF INFINITESIMAL TRANSFORMATION GROUPS TO THE SOLUTION OF NONLINEAR PARTIAL DIFFERENTIAL EQUATION
Dr. George Ullrich, US Army Mobility Equipment Research and Development Center, Ft. Belvoir, Virginia

ON UNIQUENESS IN PIECEWISE POLYNOMIAL APPROXIMATION
C. K. Chui, P. W. Smith, and J. D. Ward, Texas A & M University

LINEAR GENERALIZATIONS OF THE GRONWALL-REID-BELLMAN LEMMA
Dr. Jagdish Chandra and Paul Wm. Davis, US Army Research Office, Durham, North Carolina

A METHOD FOR THE SOLUTION OF TWO-POINT BOUNDARY VALUE PROBLEMS BASED ON THE USE OF VOLTERRA INTEGRAL OPERATIONS
Professor Louis B. Rall, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin and H. Fujita, University of Tokyo, Japan

1300-1500 TECHNICAL SESSION IV - TROPHY ROOM

CHAIRMAN: To be announced later

THE INFLUENCE OF STACKING SEQUENCE ON THE DYNAMIC BEHAVIOR OF COMPOSITE STRUCTURES
Tien-Yu Tsui, US Army Materials and Mechanics Research Centers, Watertown, Massachusetts

STRESS FIELDS AROUND PRE-CUT DAMAGE IN STIFFENED COMPOSITE PANELS
Chatta Lakshmikantham, US Army Materials and Mechanics Research Center, Watertown, Massachusetts

Wednesday PM

1300-1500 TECHNICAL SESSION IV (Continued)

NONLINEAR VIBRATION THEORY OF PAVEMENTS
Richard A. Weiss, US Army Engineer Waterways Experiment Station, Vicksburg, Mississippi

APPLICATION OF THE THEORY OF SLENDER CURVED RODS TO THE ANALYSIS OF ELASTIC YARNS
N. C. Huang, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

1500-1520 BREAK

1520-1620 GENERAL SESSION II - ERICA ROOM

SPEAKER: Professor Donald. Cohen *
Division of Mathematics
California Institute of Technology
1201 E. California Boulevard
Pasadena, California 91109

TITLE: Nonlinear Problems in Chemically Reacting Diffusive Systems

CHAIRMAN: Dr. J. Barkley Rosser
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, Wisconsin 53706

****

Thursday, 15 May 1975

0830-1000 TECHNICAL SESSION V - ERICA ROOM

CHAIRMAN: Dr. M. A. Hussain
Maggs Research Center
Watervliet Arsenal
Watervliet, New York 12189

*Dr. Viterbi spoke in General Session II.

Thursday AM

0830-1000 TECHNICAL SESSION V (Continued)

HEAT-BALANCE METHODS IN MELTING PROBLEMS
Professor Ben Noble, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

THE EFFECT OF VARIABLE DENSITY ON TRANSIENT SHEAR FORCES IN GUN BARRELS
Rao V. S. Yalamanchili, General Thomas J. Rodman Laboratory, Rock Island Arsenal, Rock Island, Illinois

NONLINEAR PROBLEMS IN THE INTERACTION OF A STEEL CARTRIDGE CASE AND CHAMBER
J. J. Toal and S. C. Chu, General Thomas J. Rodman Laboratory, Rock Island Arsenal, Rock Island, Illinois

NINE-POINT DIFFERENCE SOLUTIONS FOR POISSON'S EQUATION
Professor J. Barkley Rosser, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

0830-1000 TECHNICAL SESSION VI - TURQUOISE ROOM

CHAIRMAN: Commander
US Army Material Command
ATTN: AMCRD-TV
Mr. Herbert Cohen
Washington, D. C. 20315

ILLUMINATING ROUND EFFECTIVENESS MODELING
Dr. Martin Messinger and L. Oleniczak, Picatinny Arsenal, Dover, New Jersey

PROBABILITY OF SIGNAL SYNCHRONIZATION TIMES
Jacob Benson, Communications/Automatic Data Processing Laboratory, US Army Electronics Command, Ft. Monmouth, New Jersey

LAUNCHING OF ELECTROMAGNETIC SURFACE WAVES AT A PLANAR METAL-AIR INTERFACE
J. M. Zavada and E. L. Church, Frankford Arsenal, Philadelphia, Pennsylvania

Thursday AM

0830-1000 TECHNICAL SESSION VI (Continued)

DETERMINATION OF PROPAGATION CONSTANTS IN SCATTERING FROM DIELECTRIC-COATED WIRES
Dr. Leon Kotin, Communications/Automatic Data Processing Laboratory, US Army Electronics Command, Ft. Monmouth, New Jersey

1000-1020 BREAK

1020-1145 TECHNICAL SESSION VII - ERICA ROOM

CHAIRMAN: Dr. Clyde A. Morrison
Branch 320
Harry Diamond Laboratories
2800 Powder Mill Road
Adelphi, Maryland 20783

COMPARISON OF PERTURBATION-THEORETIC AND EXACT CALCULATIONS OF NONLINEAR OPTICAL PROPERTIES OF OPTO-ELECTRONIC MATERIALS
L. M. Narducci and R. A. Shatas, Physical Sciences Directorate, US Army Missile Command, Redstone Arsenal, Alabama, S. S. Mitra, Department of Electrical Engineering, University of Rhode Island, Kingston, Rhode Island

A TIME-DEPENDENT QUANTIZED NATURAL COLLISION COORDINATES METHOD
Norman M. Witriol, Physical Sciences Directorate, US Army Missile Command, Redstone Arsenal, Alabama

CANONICAL TRANSFORMATIONS OF THE SCHRÖDINGER EQUATION AND THE COHERENT STATE REPRESENTATION
Charles M. Bowden, Physical Sciences Directorate, US Army Missile Command, Redstone Arsenal, Alabama

SIMULATION OF THE INTERMEDIATE BALLISTIC ENVIRONMENT OF A SMALL ARM
Csaba K. Zoltani, Applied Mathematics and Sciences Laboratory, US Army Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland

Thursday AM & PM

1020-1145 TECHNICAL SESSION VIII - TURQUOISE ROOM

CHAIRMAN: Mr. A. S. Elder
Interior Ballistics Laboratory
US Army Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland 21005

GENERALIZED SHOCK WAVE PHYSICS-ELECTROMAGNETIC AND SECOND SOUND SHOCKS
Paul Harris, Picatinny Arsenal, Dover, New Jersey

ON RIEMANN'S INVARIANT AND SHOCK IMPEDANCE OF SOLIDS
Y. K. Huang, Watervliet Arsenal, Watervliet, New York

FREQUENCY DEPENDENT WAVE ARRIVAL TIME DELAYS IN DISPERSIVE MEDIA
J. R. Stabler, E. A. Baylot, and D. H. Cress, Mobility and Environmental Systems Laboratory, US Army Engineer Waterways Experiment Station, Vicksburg, Mississippi

INBORE MOTION OF ARTILLERY SHELLS
Evans H. Walker, Materiels Application Group, Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland

1145-1300 LUNCH

1300-1400 Travel by bus from Rodeway Inn to building 300; White Sands Missile Range, New Mexico

1400-1445 Briefing on National Range Operations - Conference Room Building 300.

1445-1455 BREAK

1455-1555 Observation of missile firings or tour of Nuclear Effects Laboratory, Solar Furnace facilities (depending on range schedule)

1555-1700 Travel by bus from White Sands Missile Range to Rodeway Inn

Friday, 16 May 1975

0800-0940 TECHNICAL SESSION IX - ERICA ROOM

CHAIRMAN: Director
US Army Air Mobility R & D Laboratory
ATTN: SAVDL-AS (Dr. John D. Hwang)
NASA Ames Research Center, Mail Stop 207-5
Moffett Field, California 94035

THE METHOD OF PARABOLIC SUBSTITUTION FOR HIGH SUBSONIC FLOW
Robert E. Singleton and K. Oswatitsch, US Army Research Office, Durham, North Carolina

THE BACKWARD BEAM EQUATION AND THE NUMERICAL COMPUTATION OF DISSIPATIVE EQUATIONS BACKWARDS IN TIME
Alfred Carasso, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

SPECIAL SOLUTIONS OF THE ONE-DIMENSIONAL PARABOLIC EQUATION
Siegfried H. Lehnigk, Physical Sciences Directorate, US Army Missile Command, Redstone Arsenal, Alabama

INTEGRATION OF $\int_0^\infty F(x)\, J_0(ax)\, J_1(bx)\, dx$
Sunsuke Takagi, US Army Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire

SINGULAR PERTURBATION ANALYSIS IN DIFFUSION AND HEAT CONDUCTION PROBLEMS
John F. Polk, Applied Mathematics and Sciences Laboratory, Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland

0800-0940 TECHNICAL SESSION X - TROPHY ROOM

CHAIRMAN: Patrick J. Higgins
National Range Operations Directorate
White Sands Missile Range, New Mexico 88002

SAMPLE SIZES FOR IN-FLIGHT RELIABILITY DETERMINATION
E. F. Southworth, Army Missile Test and Evaluation Directorate, White Sands Missile Range, New Mexico

Friday AM

0800-0945 TECHNICAL SESSION X (Continued)

DOVAP BEST ESTIMATE OF TRAJECTORY
R. H. Turner and W. S. Agee, National Range Operations Directorate, White Sands Missile Range, New Mexico

DOVAP INSTRUMENTATION PLANNING
W. S. Agee and J. L. Meyer, National Range Operations Directorate, White Sands Missile Range, New Mexico

PROVING PROGRAMS CORRECT
Elwood D. Baas, Army Missile Test and Evaluation Directorate, White Sands Missile Range, New Mexico

PLASTIC FLOW IN A HOLLOW PRESSURIZED CYLINDER UNDER GENERALIZED PLANE STRAIN CONDITIONS WITH REFERENCE TO THE AUTOFRETTAGE PROCESS
A. S. Elder, Interior Ballistics Laboratory, Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland

0940-100 BREAK

1000-1100 GENERAL SESSION III - ERICA ROOM

SPEAKER: Dr. Andrew J. Viterbi *
LINKABIT
10453 Roselle Street
San Diego, California 92121

TITLE: A Maximum Likelihood Decision Algorithm with Multiple Applications in Digital Communications

CHAIRMAN: Commander
US Army Electronics Command
ATTN: AMSEL-NL-H-2 (Dr. Walter Pressman)
Fort Monmouth, New Jersey 07703

1145 ADJOURN

* DR. Cohen spoke in General Session III.

ADDENDUM

Due to unforseen conflicts at the time of publication of the original program, the following changes have been made.

GENERAL SESSION II - ERICA ROOM

SPEAKER: Dr. Andrew J. Viterbi
LINKABIT
10453 Roselle Street
San Diego, California 92121

TITLE: A Maximum Likelihood Decision Algorithm with Multiple Applications in Digital Communications

GENERAL SESSION III - ERICA ROOM

SPEAKER: Professor Donald S. Cohen
Division of Mathematics
California Institute of Technology
1201 E. California Boulevard
Pasadena, California 91109

TITLE: Nonlinear Problems in Chemically Reacting Diffusive Systems

ON SPECTRALLY BOUNDED SIGNED GRAPHS*

A. J. Hoffman
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT: This is a summary of part of the lecture.

We consider two questions concerning symmetric (0,1,-1) matrices A with 0 diagonal and least eigenvalues $\lambda^1(A)$: (1) can we determine (roughly) $|\lambda^1(A)|$ by the largest order of principal submatrices of certain types? (2) can we approximate $A-\lambda^1(A)I$ by the grammian of a (0,1,-1) matrix?

We shall consider the class of all signed graphs G(i.e., graphs in which any pair of adjacent vertices are joined by a positive edge or negative edge), and their associated adjacency matrices A = A(G) defined by

$$a_{ij} = \begin{cases} 0 \text{ if } i \text{ and } j \text{ are non-adjacent vertices (thus } a_{ii} = 0 \text{ for all } i) \\ 1 \text{ if } i \text{ and } j \text{ are joined by a positive edge} \\ -1 \text{ if } i \text{ and } j \text{ are joined by a negative edge} \end{cases}$$

If A_1 is any symmetric matrix, we denote its eigenvalues in ascending order by $\lambda^1(A) \le \lambda^2(A) \le \ldots$. If G and H are signed graphs, we say $G \subset H$ (G is an induced subgraph of H) if A(G) is a principal submatrix of H.

The first question we consider, in analogy to [3] for ordinary graphs, is whether the magnitude of $\lambda^1(G) \equiv \lambda^1(A(G))$ can be in anyway related to the size of certain induced subgraphs of G. To explain the question (and the answer), we must first introduce some notation:

A large empty circle ○ will denote an independent set of t vertices. A large circle with a + inside ⊕ will denote a set of t vertices, every pair of which are joined by a positive edge; a large circle with a - inside ⊖ will denote a set of t vertices, every pair of which is joined by a negative edge. The symbol

t: $\overset{+}{\_\_\_\_}$?

means that a single vertex is joined by a positive edge to each of t vertices, whose relation to each other is specified by ? (All + all - or independent). The meaning of

T: $\overset{-}{\_\_\_\_}$?

should be clear.

* This work was supported (in part) by the Army Research Office under contract number DAHC04-74-C-0007.

Finally,

t: ⓣ $\overset{+}{\text{———}}$ ⓣ

means that each of the left hand set of t vertices is joined by a positive edge to each of the right hand set of t vertices; and the symbol

t: ⓣ $\overset{-}{\text{———}}$ ⓣ is similarly defined.

Consider now the following graphs:

(1) t: ⊖

(2) t: $\overset{+}{\text{———}}$ ○

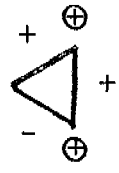
(3) t: $\overset{-}{\text{———}}$ ○

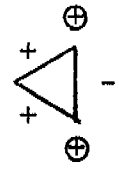
(4) t: $\overset{+}{\text{——}}$ ⊕ $\overset{+}{\text{——}}$ ⊕

(5) t: $\overset{+}{\text{——}}$ ⊕ $\overset{-}{\text{——}}$ ⊕

(6) t: $\overset{-}{\text{——}}$ ⊕ $\overset{+}{\text{——}}$ ⊕

(7) t: $\overset{-}{\text{——}}$ ⊕ $\overset{-}{\text{——}}$ ⊕

(8) t: 

(9) t: 

(10) t: 

We define, for each graph G, t(G) to be the smallest integer t such that none of (1)-(10) is an induced subgraph of G.

One copy can readily prove that t(G) cannot be large if $|\lambda^1(G)|$ is of modest size. The reason is that, if H_t is any one of (1)-(10), $\lambda^1(H_t)\to-\infty$ as $t\to\infty$. But $H_t\subset G$ implies $\lambda^1(H_t)\geq\lambda^1(G)$. The point of interest is the converse. If t(G) is modest, so is $|\lambda^1(G)|$ in the following sense:
Theorem 1. Each of the two functions of signed graphs, t(G) and $|\lambda^1(G)|$ is bounded by a function of the other.

We will not be specific about these functions (here and in the theorems stated below), since the functions produced by the proofs (which depend heavily on the use of Ramsey's theorem and its relatives [5]) are enormous, and manifestly very far from being accurate. The points to be emphasized are

(i) the functions are functions of one variable only, and in particular do not depend on the number of vertices; and

(ii) it would be of interest to find better estimates of these functions, closer to the truth.

The second question to be considered, inspired by [1], is the following. Let G be a signed graph, then how can we represent the positive semi-definite matrix $A(G)-\lambda^1(G)I$ by a grammian KK^T? In case $\lambda^1(G)=-2$, the question has a nice geometric interpretation. The rows of K are vectors of length $\sqrt{2}$, and the inner products of the rows of K are 0,1,-1. Hence, the question is equivalent to finding ways of placing vectors in Euclidean space so that any two make an angle of 0^o,60^o, or 120^o. From the theory of root systems in Lie groups [2], one can infer that, if G is connected, then there exists a (0,1,-1) K satisfying

$$A(G)-\lambda^1(G)I = KK^T \tag{11}$$

with finitely many exceptions.

We are led to conjecture that (11) might hold for other integral values of $\lambda^1(G)$.

Conjecture 1. If $\lambda^1(G)$ is integral and G is connected, then, with finitely many exceptions there exists a (1,-1,0) matrix K such that (11) holds.

It is easy [6] to show that Conjecture 1 is true (with no exceptions) if $\lambda^1(G) = -1$. But conjecture 1 is false for all other integral $\lambda^1(G) = -3,-4,\ldots$. The counter-examples are constructed inductively as follows. Find a G such that $\lambda^1(G) = -r$, and (11) does not hold for that G(this can be done for r = 2). Make up a graph H consisting of an enormous number of copies of G (enormous depends on r), with all instances of vertex

i of G (for all i) joined by a positive edge. Then (11) will be false for H, and $\lambda^1(H) = -(r+1)$. There will be an infinite number of such H, since we can take an arbitrarily large number of copies of G to make H. Now the induction continues, using only one of the H in place of G.

There is another reasonable conjecture (whether or not $\lambda^1(G)$ is an integer):

Conjecture 2. There exists a constant C such that, for each signed graph G, there exists a (1,-1,0) matrix K with

$$||(A(G) - \lambda^1(G)I) - KK^T|| \leq C. \tag{12}$$

Here, $||\ ||$ stands for the usual operator norm.

Conjecture 2 is also false, by the following example. Assume C given. Choose an integer $m > C + \frac{1}{2}$. Choose an integer n so that

$$\sqrt{4m^2 + 4n^2} - 2n < 1, \text{ and}$$

$$n - \frac{n^2 - C}{m+n+C} < \frac{1}{2} .$$

Now construct a graph $G^{(p)}$ as follows. Begin with a claw K_{1,n^2}. Denote the n^2 vertices of the claw by $1,\ldots,n^2$. In addition, construct 2m (positive) cliques, each of size p, adjacent to 1; 2m positive cliques, each of size p, adjacent to 2;...; 2m positive cliques, each of size p, adjacent to n^2. It can be shown that, for p large enough, (12) will be false.

It turns out, however, that a weak version of (12) is true, with C depending on λ^1. In fact, we have the following results, improving [3], and using methods of [3] and [4].

Theorem 2. There exists a fixed function f(x) such that, for every signed graph G, there exists a (+1,-1,0) matrix K such that

$$||(A(G) - \lambda^1(G)I) - KK^T|| \leq f(\lambda^1(A)). \tag{13}$$

Further, the number of nonzero entries in each row of K is

$$\min([-\lambda^1(G), -\lambda^1(\text{abs } G)]).$$

(Here, [x] is the greatest integer at most x, abs G is the ordinary graph obtained from G by making all edges positive).

Theorem 3. There exists a fixed function f(x) such that, for every signed graph G, there exists a (+1,-1,0) matrix K, every row of K

containing min ($[-\lambda^1(G)]$, $[-\lambda^1(\text{abs } G)]$) nonzero entries, so that (13) and

(14) $$|| A(\text{abs } G) - \lambda^1(\text{abs } G)I - (\text{abs } K)\ (\text{abs } K)^T|| \leq f(\lambda^1(A))$$

hold. (Here, abs K is obtained from K by replacing each -1 by +1).

REFERENCES

[1] Cameron, P. J., Goethals, J. M., Seidel, J. J., and Shult, E. E., "Line graphs, root systems, and elliptic geometry", to appear.

[2] Carter, R. W., "Simple groups of Lie type", Wiley (1972).

[3] Hoffman, A. J., "On spectrally bounded graphs", in A Survey of Combinatorial Theory, edited by J. N. Srivatava, North-Holland, 1973, 277-283.

[4] Hoffman, A. J., "On eigenvalues of symmetric (+1,-1) matrices", Israel Journal of Mathematics 7(1974), 69-75.

[5] Hoffman, A. J., "Applications of Ramsey style theorems to eigenvalues of graphs", in Combinatorics, Part 2, edited by M. Hall, Jr. and J. H. van Lint, Mathematical Centre Tracts 56, Mathematisch Centrum, Amsterdam, 1974, 43-58.

[6] Hoffman, A. J., and Francisco Pereira, "On copositive matrices with (-1,0,1) entries", J. Comb. Theory A14 (1973), 302-309.

EFFECTS OF CONCENTRATED MASS AND THRUST DIRECTIONAL CONTROL ON THE STABILITY OF FREE-FREE BEAMS

Julian J. Wu
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. The application of the finite element technique to non self-adjoint, more-than-two-points boundary value problems was demonstrated in a recent paper on missiles' stability analysis. Numerical results are limited there. However, they have brought new understanding on the basic stability behaviors of a free-free beam. The effect on such behaviors due to a concentrated mass, modeling a piece of heavy machinery, was also shown for some special cases.

The present paper is an extension of the previous work to include the location as well as the amount of the concentrated mass as parameters. Also included is an extensive study of the thrust directional control feedback and its effect on the stability behavior's of free-free beams.

1. INTRODUCTION. The study of nonconservative stability of structures without fixed supports* has turned out many surprises. Although extremely similar to those with fixed supports in terms of the governing equations, the so-called "free" structures present difficulties not only in methods of solutions but in the interpretation of the solutions themselves. Many basic features of the problem were not sufficiently exploited in the literature dealing with "free" structures [1-8]. As Solarz pointed out in conjunction with a "free-free" beam problem [6], that the assumption of a fixed end could remove some important features of the structural motion as a "rigid body" with its vibrations. No adequate explanations were given, however. And to the best of our information, none was found elsewhere.

Let us ask a trivial question, for example: How many zero eigenvalues are there associated with the vibration of a uniform, free-free beam? The answer is two; and this can be easily shown by going through the process of the separation of variables for partial differential equations. However, in some of the most well-known text (i.e. [9, 10]), only one zero eigenvalue was reported. Probably such an inaccuracy is of no consequence for the cited example per se, except for the case of the same beam but now

*As an initial step, only beam structures are considered since they are the simplest possible continuous systems.

also subjected to an initial axial force. Since then, the realization of the correct number of zero eigenvalues of the first problem is essential to the proper interpretation of the stability data of the second. A numerical analysis of such problems and some other points have already reported by the present writer [11, 12].

The purpose of the present paper is to include some additional computational results on the stability behavior of a free-free beam. In particular, the effect due to the amount and location of a concentrated mass and the amount and the sign of the directional control parameter are analyzed.

2. PROBLEM FORMULATION. Let us consider the plane motion of a uniform missile attached with a concentrated mass travelling under a constant acceleration in the axial direction. The only forces are the thrust P and the inertia force $(\rho A\ell+M)a$, where ρ = density of the material, A = area of cross-section, ℓ = length of the missile, M = amount of the concentrated mass and a = axial acceleration.

Using the notion well described by Bolotin [13], we shall first discuss the undisturbed form of equilibrium. The stability of such an equilibrium will then be determined by an investigation of the disturbed motion, which is the prime concern of our analysis.

The undisturbed state (Figure 1) is governed by three equations:

(1a) $\Sigma F_X = 0$

(1b) $\Sigma F_Y = 0$

(1c) $\Sigma M = 0$

Equations (1) states that the total force in X-direction, in Y-direction and the total moment in X-Y plane are zero. From equation (1a), we have

(2a) $P = (\rho A\ell+M)a$

or

(2b) $a = \frac{P}{\rho A\ell+M}$

Thus the initial force acting on the cross-section of the beam is given by

$$(3) \qquad T(x) = \begin{cases} (\rho Ax)a = \frac{x}{\ell}\left(\frac{P}{1+m}\right), & 0 \le x < x_m \\ (\rho Ax+M)a = \left(\frac{x}{\ell} + m\right)\left(\frac{P}{1+m}\right), & x_m \le x < \ell \end{cases}$$

where x is measured from the tip of the beam, $m = \frac{M}{\rho A \ell}$ and x_m denotes the location of the concentrated mass. Equation (1b) is always satisfied because there is no force acting in Y direction. So is equation (1c) due to the fact that the thrust and the inertia force are colinear in the undisturbed state.

The disturbed motion will be defined by a small lateral displacement u(x,t), perpendicular to the undisturbed (rectilinear) axis (Figure 2). As a means to manipulate the stability behavior, we assume that the direction of P can be rotated through a small angle θ about the tangent at the tail end of the disturbed beams such that

$$(4) \qquad \theta = K_\theta u'(x_\theta, t)$$

where K_θ is a nondimensional design constant, x_θ denotes the location of the direction control sensor. The only limitation on K_θ is that θ be small so that the linearized equations of disturbance are valid. These equations consist of the following:

$$(5a) \quad \text{D. E.} \qquad EIu'''' + [T(x)u']' + \rho A\ddot{u} + M\ddot{u}(x_m)\,\delta(x-y_m) = 0$$

$$(5b,c) \quad \text{B. C.} \qquad u''(0) = 0, \qquad u'''(0) = 0$$

$$(5d) \qquad u''(\ell) = 0$$

$$(5e) \qquad EIu'''(\ell) = P\theta = PK_\theta u'(x_\theta)$$

where E = Young's modulus, I = second moment of the cross-section and $\delta(x)$ is the Dirac function. T(x) is given by equation (3) and θ, by (4). A prime (') denotes a differentiation with respect to the spatial coordinate x and a dot ($\cdot$), a differentiation with respect to time t. Equations (5) can be conveniently derived by integrating the three-dimensional equations for the disturbed motion given in reference [14].

It will be convenient to write equations (5) in terms of dimensionless quantities. Thus,

$$(6a) \quad \text{D.E.} \qquad u'''' + (f(x)u')' + \lambda^2 u + m\lambda^2 u(x_m)\,\delta(x-x_m) = 0$$

$$(6b,c) \quad \text{B.C.} \qquad u''(0) = 0, \qquad u'''(0) = 0$$

(6d) $$u''(1) = 0$$

(6e) $$u'''(1) = K_\theta Q u'(x_\theta) = 0$$

where f(x) in (6a) is given by

(7) $$f(x) = \begin{cases} Q_1 x, & 0 \le x \le x_m \\ Q_1 x + m Q_1, & x_m \le x \le 1 \end{cases}$$

$Q_1 = Q/(1+m)$ and $Q = P\ell^2/(EI)$.

The spatial variables are made dimensionless through a division by ℓ. The time t is made dimensionless through a division by the constant $c = (\rho A \ell^4/EI)^{1/2}$, which is in real time unit (seconds, for example). In equations (6), we have also assumed that the solution of u(x,t) is exponential in time, i.e.,

(8) $$u(x,t) = u(x)e^{\lambda t}.$$

There should be no confusion from using the same letter u for two different functions u(x,t) and u(x).

3. BASIS OF SOLUTIONS AND THE ADJOINT PROBLEM. Numerical solutions to the equations of the disturbed motion (equations (6)) will be obtained through the finite element technique - adjoint variational formulations. The basis of this formulation is given here. Through integrations-by-parts, it is straightforward to show that the adjoint problem of equations (6) and the associated variational statement are given in equations (9) and (10) respectively.

(9a) D.E. $$v'''' + [f(x)v']' + \lambda^2 v + m\lambda^2 v(x_m)\,\delta(x-x_m) - QK_\theta\, v(1)\,\delta'(x-x_\theta) = 0$$

(9b,c) B.C. $$v''(0) = 0, \qquad v'''(0) = 0$$

(9d) $$v''(1) + Qv(1) = 0$$

(9e) $$v'''(1) + Q\,v'(1) = 0$$

and

(10a) $$\delta I = 0$$

(10b) $$I = \int_0^1 [u''v'' - f(x)u'v' + \lambda^2 uv + m\lambda^2 u(x_m)v(x_m)\delta(x-x_m)]dx + Qu'(1)\delta v(1) + QK_\theta u'(x_\theta)v(1)$$

where $v(x)$ is the adjoint field variable and $f(x)$ is given in Eq. (7).

It is worthwhile to note that the boundary conditions of Eqs. (6) and (9) are all natural boundary conditions. Thus the variational statement (10) is completely unconstrained. In other words, the coordinate functions used for Rayleigh-Ritz approximation need not to satisfy any of the boundary conditions [15].

We further note that Eq. (6) is not exactly a two point boundary value problem unless $x_\theta = 1$. It is clear that the boundary condition of the type of Eq. (6e) in which $x_\theta \neq 1$ does not present any difficulty when the variational statement (10) is used for finite element solutions.

For the special case $x_\theta = 1$, the adjoint problem of Eqs. (9) can be written in a slightly different form:

(11a) D.E. $$v''''+[f(x)v']'+\lambda^2 v+m\lambda^2 v(x_m)\delta(x-x_m) = 0$$

(11b,c) B.C. $$v''(0) = v'''(0) = 0$$

(11d) $$v''(1) + Q(1+K_\theta)\, v(1) = 0$$

(11e) $$v'''(1) + Qv'(1) = 0$$

Once the variational statement has been established, it is a routine matter to apply the finite element discretization. In conjunction with adjoint variational statements, this procedure has also been well documented [11, 15, 16, 17] and will not be repeated here.

4. STABILITY BEHAVIOR DUE TO DIRECTIONAL CONTROL. Prior to the presentation of numerical results and discussions, it would be worthwhile to recapitulate some basic definitions and concepts. As mentioned in Section 2, the disturbance is assumed to be in the form of

(10) $$u(x,t) = u(x)e^{\lambda t}.$$

The eigenvalue λ is a complex number in general. Thus we write

(11) $$\lambda = \lambda_R + i\,\lambda_I$$

where $i = \sqrt{-1}$. Both λ_R and λ_I are real numbers. When λ_R is negative or zero, the disturbance $u(x,t)$ will decrease with time or remain finite, and the structure is thus considered stable. When λ_R is nonzero and positive, $u(x,t)$ will grow with time and instability occurs. Divergence instability is characterized by $\lambda_I = 0$ and flutter (or oscillatory) instability, by $\lambda_I \neq 0$. In the particular case when both λ_R and λ_I are zero, the solution represents a rigid body motion. According to the definition of stability described in reference [13], it implies that small disturbance leads to small deviation from the unidsturbed state. In case of small rigid body motion, disturbance and deviation are the same. Therefore the rigid body

translation and rotation associated with the small disturbance must be considered stable modes.

In a continuous system, there are infinite number of eigenvalues corresponding to various modes of vibrations. Since a large λ_I represents a high frequency mode which is less likely to realize than the lower frequency modes, only the lower end of the frequency spectrum is of physical importance. In a finite element discretized system, depending on the number of elements used, a finite number of eigenvalues are obtained to approximate the actual eigenvalues of lowest magnitudes. For the data presented in this paper, nine (9) elements were used. From our previous experience on similar problems, the accuracy of these data should be within about one percent (1.0% or less compared with the exact solutions [11]).

First let us consider the case without thrust directional control ($K_\theta = 0$). The numerical values of λ of the four lowest magnitudes are given in Table 1 for various values of Q and the curves of λ vs. Q are shown in Figure 3*. As shown in Table 1, the first branch of eigenvalues is zero throughout the range of Q. This solution corresponds to a rigid body translation and is well known. The second branch has eluded many investigators. At $Q = 0$, there is another zero eigenvalue which corresponds to a (small) rigid body rotation. For $Q > 0$, however, λ takes a (positive) real value. This mode of disturbance must be a bending mode because it is not a rigid body mode. Its magnitude will no longer remain small but grow with time. This unstable bending mode of disturbed motion was first realized by the present writer [11].

Again, the third and the fourth branches of eigenvalues are well known [3, 4, 8]. Prior to $Q_{CR} = 10.93\pi^2$, the eigenvalues are pure imaginary, indicating stable, oscillatory motion. At Q_{CR}, these two branches coalesce and, beyond which, the values of λ become complex, indicating flutter instability. Due to the unstable bending mode of the second branch, however, this critical thrust Q_{CR} has lost its significance and we thus refer to it as a pseudo-critical thrust (Figure 3).

For a negative K_θ ($K_\theta = -0.1$ for example), the first bending mode is associated with a positive real λ, as shown in Figure 5, indicating divergence instability. For a given positive K_θ, however, a region of stability exists between $Q = 0$ and $Q = Q_{CR}$, beyond which the structure becomes unstable due to divergence. As reported in the previous investigations [3, 8], this critical thrust Q_{CR} was thought to be a constant for all values of K_θ. This is not conceivable in the light of the present analysis. Since $Q_{CR} = 0$ at $K_\theta = 0$ due to the unstable first bending mode,

*When λ is either a purely real or a purely imaginary number, it appears as a pair of the same absolute value but with opposite signs. Only the positive ones are shown. In Figures 3, 5 and 6, we have plotted λ_I vs. Q in the upper plane and (positive) λ_R vs. Q in the lower plane.

we expect Q_{CR} to vary (continuously) from zero as K_θ increases. Again, this is substantiated by our numerical results. In Figure 6, the detailed variations of the eigenvalue of the first bending mode vs. the thrust parameter Q is shown for various values of K_θ from $K_\theta = 0$ up to 0.5. The trend of Q_{CR} from zero as K_θ increases is clearly observed. The variation of Q_{CR} vs. K_θ is again plotted in Figure 7. This variation of Q_{CR}/π^2 is shown to be extremely sensitive in the range from 0.007 to 0.05. It approaches rapidly and asymptotically to the value of 2.58 as K_θ becomes greater than 1.0. This curve was substituted by a constant value of 2.60 in the previous investigations.

Before we conclude this section, let us consider the special case of a uniform missile under a constant thrust fixed in the direction of the undisturbed axis. This problem was first considered by Silverberg* [1] and is clearly a subcase of the present analysis with $x_\theta = 1.0$ and $K_\theta = -1.0$ (i.e., $\theta = -v'(1)$). By a purely analytical approach, Silverberg was able to obtain the first non-zero Q at which the eigenvalue vanishes ($Q = 2.60\pi^2$). Thus he concluded that $Q = 2.60\pi^2$ is the buckling load. From the present analysis as shown in Figure 4, however, even though $Q = 2.55\pi^2$ is a point of vanishing eigenvalue, it cannot be considered a critical load due to the fact that the structure is unstable prior to this load. This is indicated by the real positive eigenvalue curve for $K_\theta = -1.0$ for $Q < 2.55\pi^2$ which is a region of divergence instability. Thus we have here an example showing the stability nature of a structure cannot be determined merely by seeking out the loading parameters of vanishing eigenvalues.

5. STABILITY BEHAVIOR DUE TO A CONCENTRATED MASS**. The effects on the stability behavior of a uniform beam due to a concentrated mass are shown in Figures 7 through 11.

For a tip mass of two percent of the total mass of the beam ($x_m = 0$, $m = 0.02$), there is a region of stability between $Q_1 = 2.7\pi^2$ and $Q_2 = 10.6\pi^2$. For Q less than Q_1, the beam is unstable due to divergence of the first bending mode. For Q greater than Q_2, flutter occurs due to the coalescence of the second and the third bending modes (Figure 7). As the tip mass increases ($m = 0.04$ in Figure 8 and $m = 0.06$ in Figure 9), the region of stability varies in such a manner that the values of Q_1 and Q_2 decrease ($Q_1 = 1.5\pi^2$ and $Q_2 = 8.2\pi^2$ for $m = 0.04$; $Q_1 = 0.9\pi^2$ and $Q_2 = 6.8\pi^2$ for $m = 0.06$).

It is also observed that, in Figures 8 and 9, that flutter occurs as the first and the second bending mode coalesce whereas in Figure 7, it occurs as the second and the third branches coalesce.

*Matsumoto and Mote credited the solution to Beal. But Beal himself rightfully attributed it to Silverberg.

**The data reported here supersedes some of those in Ref. [11] as an error was found in the computer program used for previous calculations.

Some indications of the effect due to the location of the concentrated mass can be seen from Figures 9, 10 and 11 in which the stability curves are calculated for a concentrated mass m = 0.06 placed at locations $x_m = 0$, $x_m = 0.5$ and $x_m = 1.0$ respectively. At locations $x_m = 0.5$ and 1.0, the curves resembles the ones with no concentrated mass (Figure 3), i.e., there is no region of stability as shown in Figure 10 and 11. However at $x_m = 0$, a region of stability can be attained within the operating thrust range of $Q_1 = 0.9\pi^2$ and $Q_2 = 6.8\pi^2$ (Figure 9).

REFERENCES

1. S. Silverberg, "The effect of longitudinal acceleration upon the natural modes of vibration of a beam," Technical Report, Space Technology Laboratories, TR-59-0000-00791, August 1959.

2. J. Kacprzynski and S. Kaliski, "Flutter of a deformable rocket in supersonic flow," Proceedings of International Aeronautical Contress, Zurick, 4, 1960, pp. 911-925.

3. T. R. Beal, "Dynamic stability of a flexible missile under constant and pulsating thrusts," AIAA Journal, 3 (3), 1965, pp. 486-494.

4. V. I. Feodosev, "On a stability problem," PMM, 29, (2), 1965, pp. 391-392 (translation from Russian).

5. S. Kaliski and S. Woroszyl, "Flutter of a deformable rocket in supersonic flow according to the second asymptotic approximation," Proceedings of Vibration Problems, 1 (6), 1965, pp. 49-81.

6. L. Solarz, "The mechanism of the loss of stability of a non-guided deformable rocket," Proceedings of Vibration Problems, 4, (10), 1969, pp. 425-441.

7. K. S. Kolesnikov and M. M. Il'in, "Dynamic stability of a uniform, controlled, beam with free ends," Izdatel'stvo Nauka, 1973, pp. 87-93 (in Russian).

8. G. Y. Matsumoto and C. D. Mote, "Time delay instability in large order systems with controlled follower forces," Journal of Dynamic Systems, Measurement and Control, Transaction, ASME, December 1972, pp. 330-334.

9. R. L. Bisplinghoff, H. Ashley and R. L. Halfman, Aeroelasticity, Addison-Wesley, Reading, Mass., 1955, p. 77-78.

10. K. N. Tong., Theory of Mechanical Vibration. John Wiley and Sons, New York, London, 1959, p. 257.

11. J. J. Wu, "Missile stability using finite elements - an unconstrained variational approach," Proceedings, 1975 Army Numerical Analysis and Computers Conference.

12. J. J. Wu, "On the stability of a free-free beam under axial thrust subjected to directional control," to appear in Journal of Sound and Vibration.

13. V. V. Bolotin, Nonconservative Problems of the Theory of Elastic Stability, MacMillan Company, New York, 1967, pp. 43-46, section 1.8.

14. Ibid, equations (1.34), (1.35) and (1.36).

15. J. J. Wu, "A unified finite element approach to column stability problems," Development in Mechanics, 8, 1975.

16. J. J. Wu, "Column instability under nonconservative forces, with internal and external damping--finite element using adjoint variational principles," Development in Mechanics, 7, 1973, pp. 501-514.

17. J. J. Wu, "On the numerical convergence of matrix eigenvalue problems due to constraint conditions," Journal of Sound and Vibration, 37 (3), 1974, pp. 349-358.

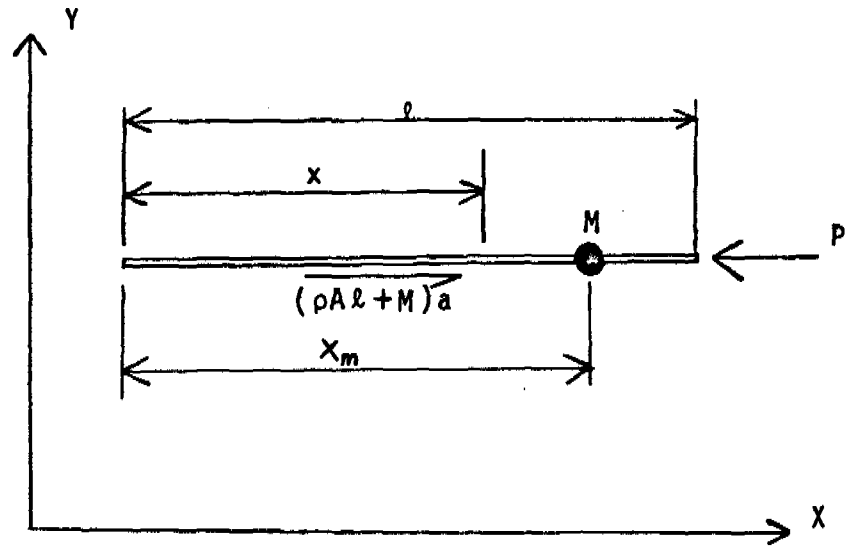


FIGURE 1. THE UNDISTURBED EQUILIBRIUM OF A FREE-FREE BEAM UNDER A CONSTANT THRUST AND THE INERTIA FORCE

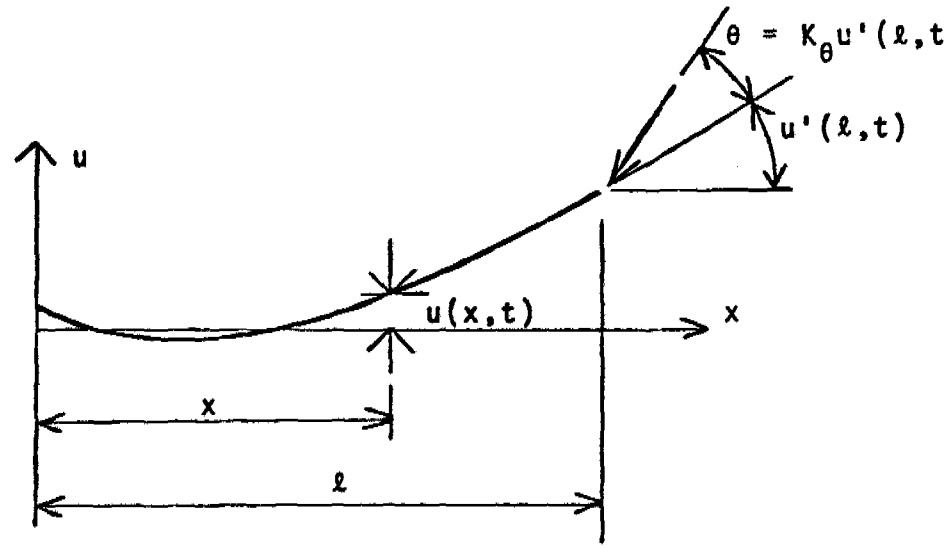


FIGURE 2. A SMALL DISTURBANCE FROM THE UNDISTURBED AXIS

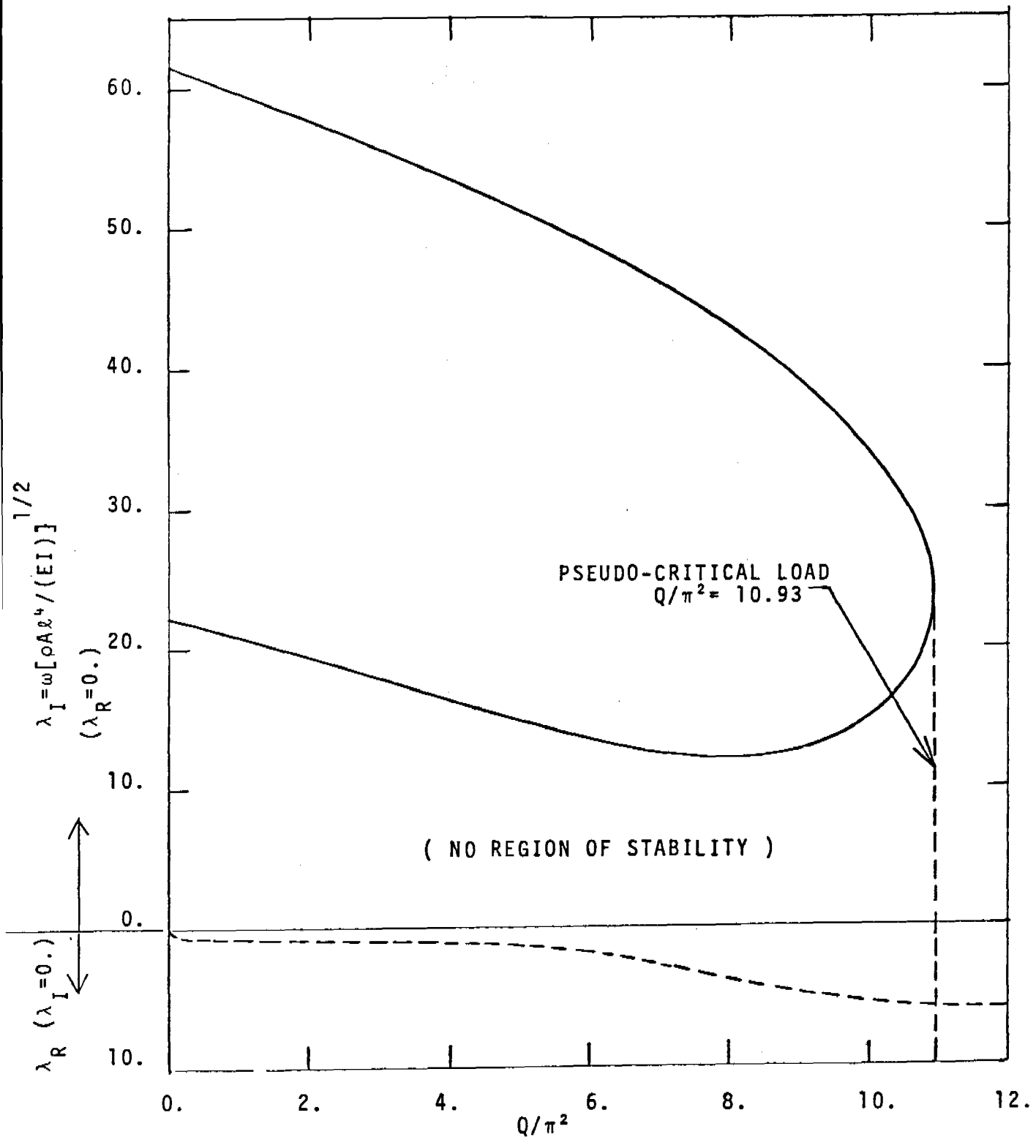


FIGURE 3. BRANCH CURVES OF THE FOUR LOWEST EIGENVALUES --- DISTURBANCE OF A FREE-FREE BEAM UNDER A CONSTANT THRUST WITHOUT CONTROL

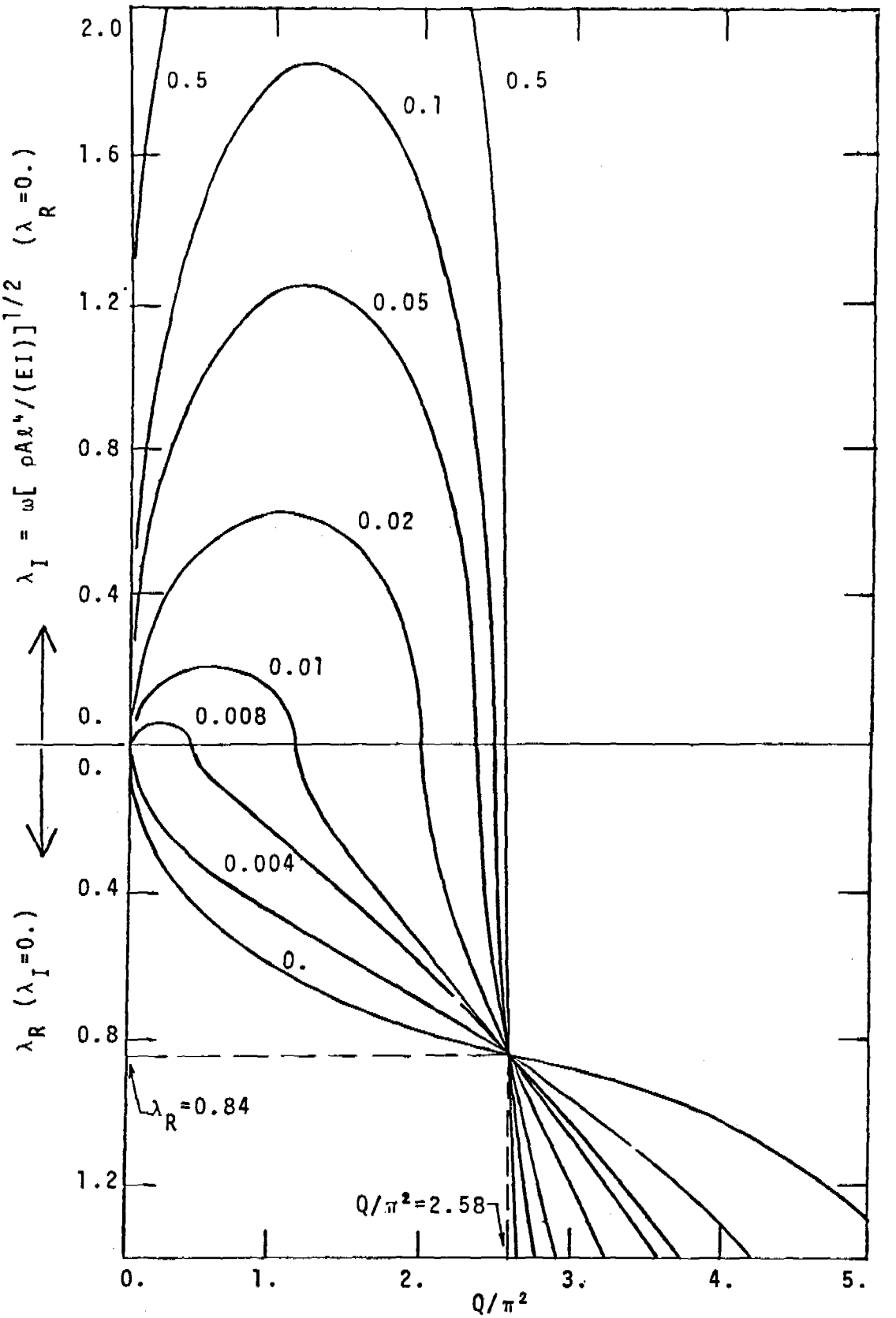


FIGURE 4. A DETAILED PLOT OF THE FIRST NONZERO EIGENVALUE CURVE FOR SMALL VALUES OF K_θ

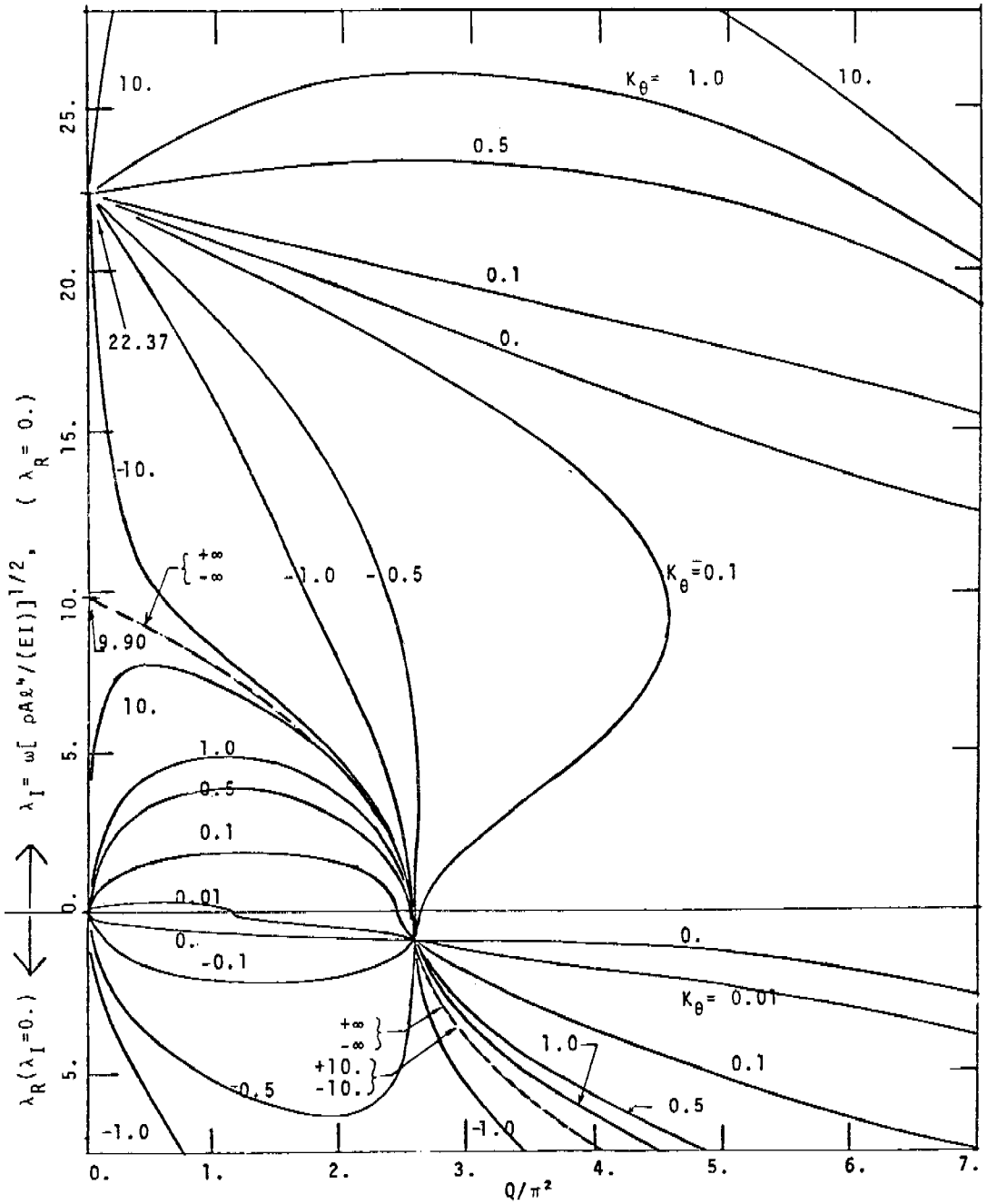


FIGURE 5. THE FIRST AND SECOND NON-ZERO EIGENVALUE CURVES FOR VARIOUS DIRECTION CONTROL PARAMETERS K_θ

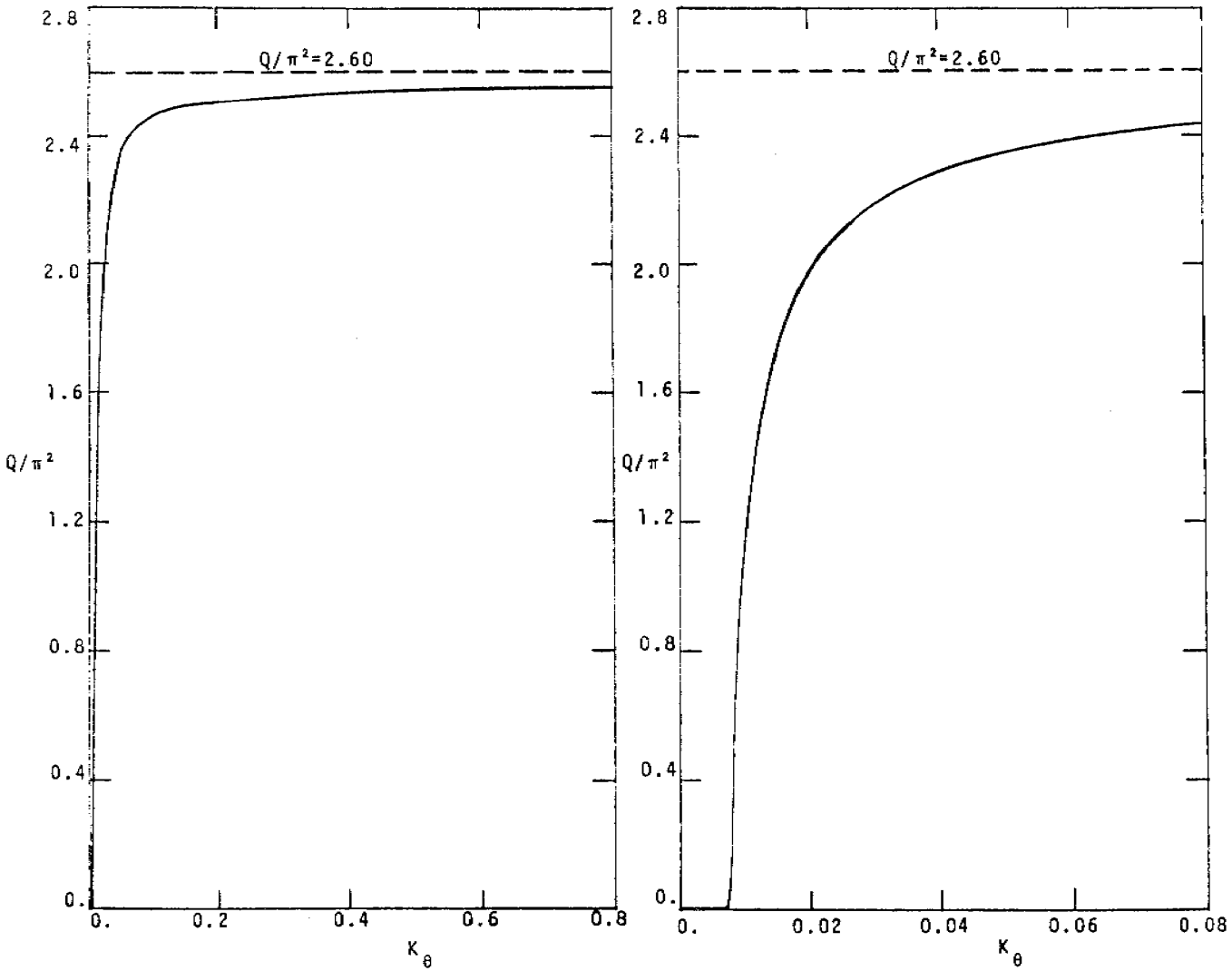


FIGURE 6. CRITICAL THRUST Q_{CR} VS. DIRECTIONAL CONTROL PARAMETER K_θ

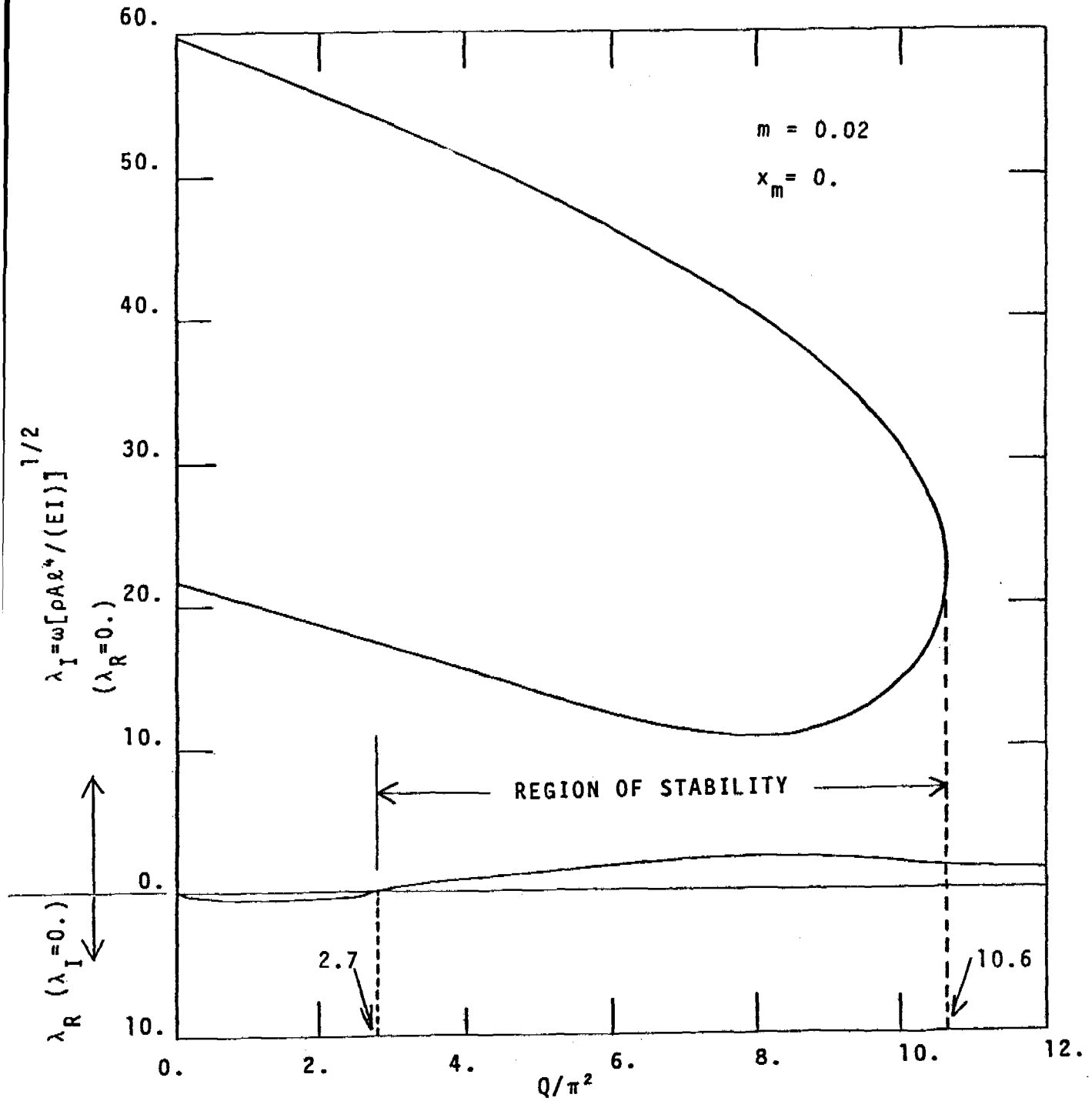


FIGURE 7. BRANCH CURVES OF THE LOWEST EIGENVALUES (x_m = 0, m = 0.02).

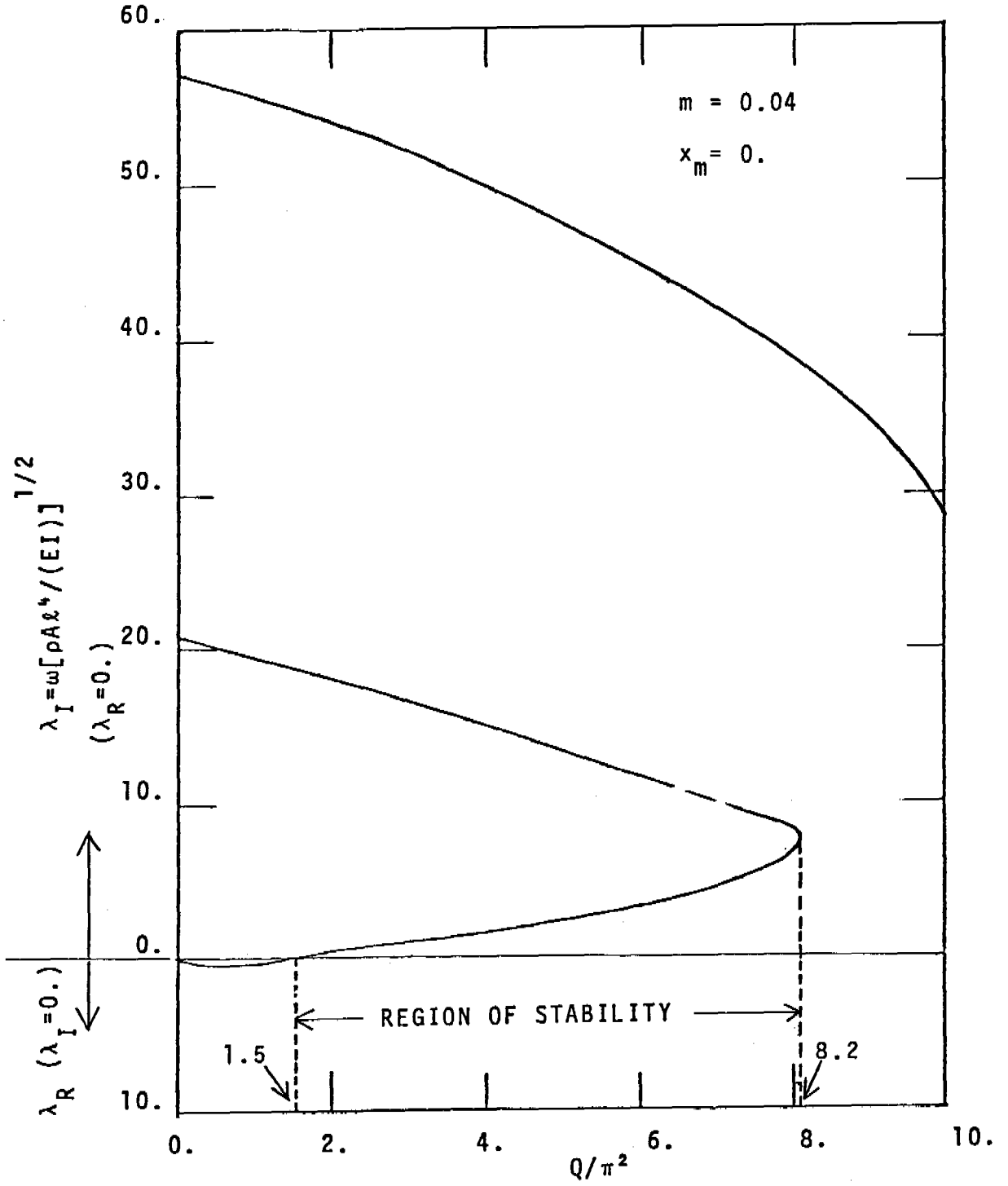


FIGURE 8. BRANCH CURVES OF THE LOWEST EIGENVALUES (x_m = 0, m = 0.04).

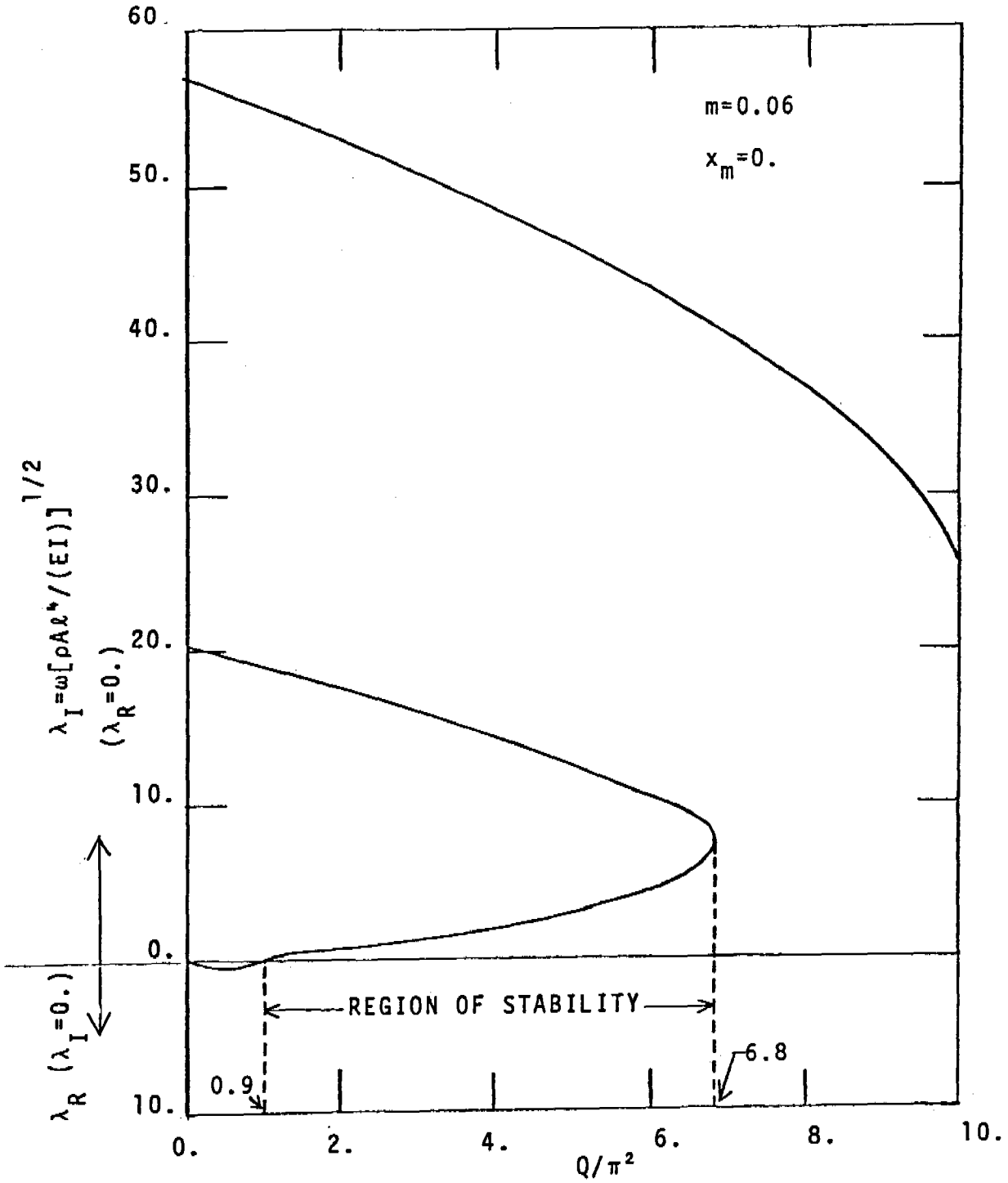


FIGURE 9. BRANCH CURVES OF THE LOWEST EIGENVALUES (x_m = 0, m = 0.06).

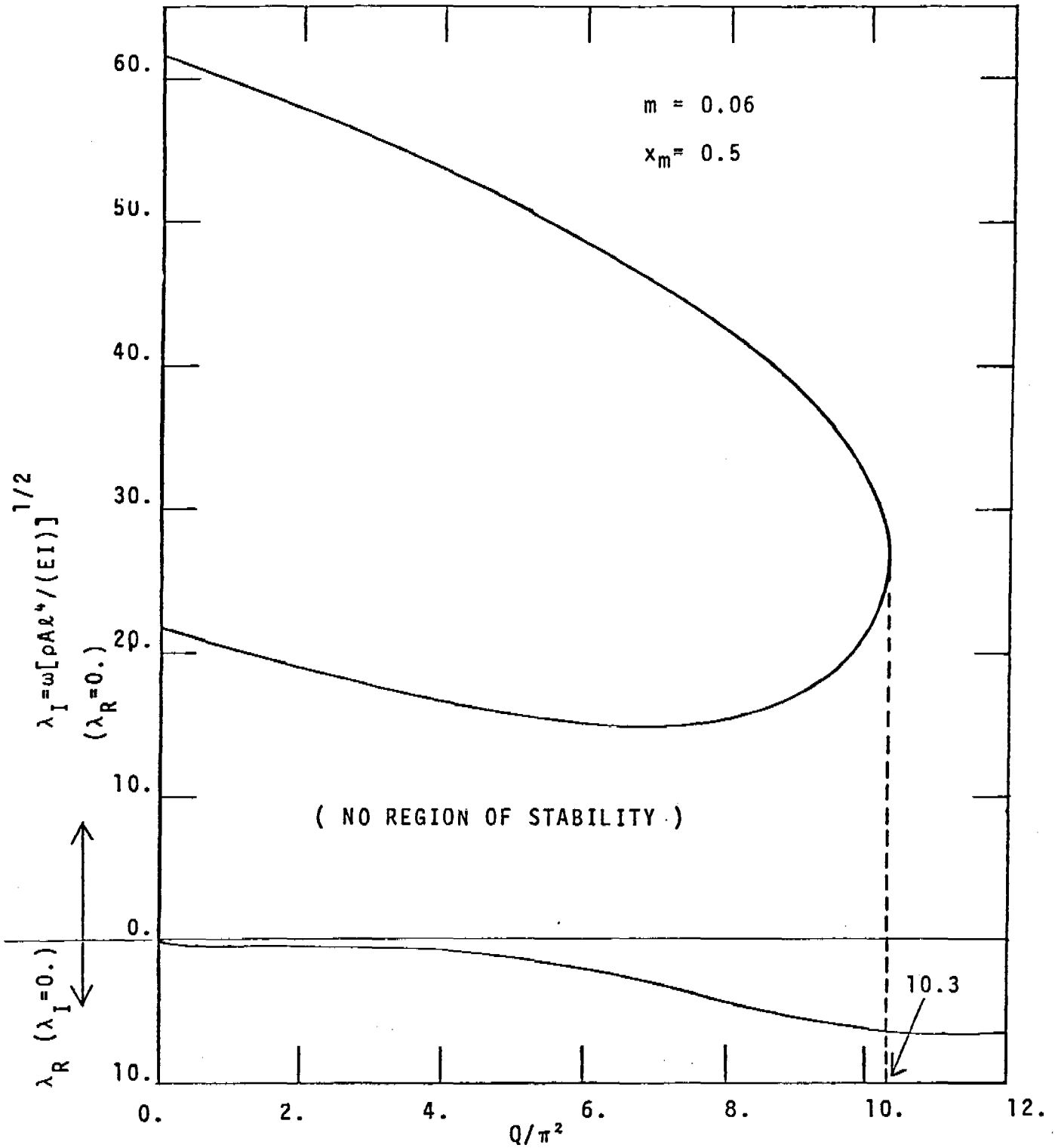


FIGURE 10. BRANCH CURVES OF THE LOWEST EIGENVALUES (x_m = 0.5, m = 0.06).

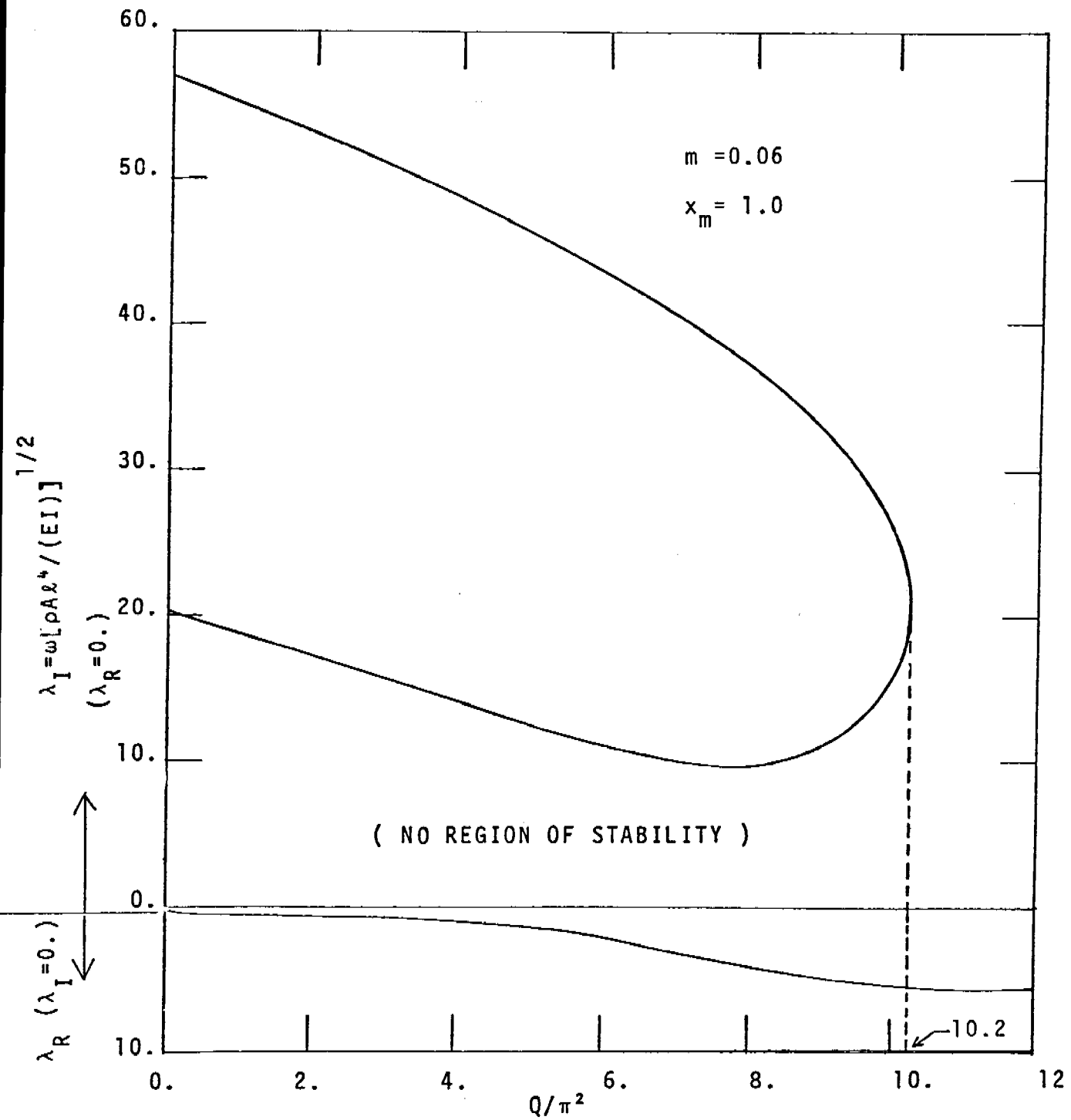


FIGURE 11. BRANCH CURVES OF THE LOWEST EIGENVALUES (x_m = 1.0, m = 0.06).

TABLE 1. NUMERICAL VALUES OF THE FOUR LOWEST EIGENVALUES FOR A UNIFORM FREE FREE BEAM UNDER A CONSTANT THRUST WITHOUT DIRECTIONAL CONTROL

Q/π^2	0.	1.	2.	3.	4.
λ_4	61.70	59.78	57.79	55.69	53.47
λ_3	22.37	20.95	19.49	17.98	16.45
λ_2	0.	(0.60)	(0.78)	(0.88)	(1.02)
λ_1	0.	0.	0.	0.	0.

Q/π^2	5.	6.	7.	8.	9.
λ_4	51.11	48.56	45.76	42.62	38.90
λ_3	14.94	13.57	12.50	12.08	12.70
λ_2	(1.29)	(1.82)	(2.64)	(3.69)	(4.72)
λ_1	0.	0.	0.	0.	0.

THE STANDARD LINEAR MODEL IN THE STABILITY AND MASS OPTIMIZATION OF NONCONSERVATIVE EULER BEAMS

Charles R. Thomas
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. A dimensionless stability problem together with its adjoint problem has been formulated for nonconservative, cantilevered Euler beams with linear external damping and internal damping according to the standard linear model. Mass optimization is considered for a beam with a linear distributed tangential load acting along the centerline. Graphical optimization plots are shown and utilized as initial guesses in a Rosenbrock optimization routine which indicates mass reductions in the range of 20% to in excess of 30% are possible.

1. INTRODUCTION. The stability problem was formulated for nonconservative, cantilevered Euler beams with linear external damping and internal damping according to the standard linear model. A convenient dimensionless form of the original equations was introduced and an adjoint system of equations was derived. These equations were then utilized in developing an adjoint variational principle which yielded a characteristic equation for critical flutter load after a proper application of a generalized Ritz procedure. Specific numerical results are then given for Hauger's problem, that is for a beam with a linear distributed tangential load acting along with the centerline of the beam.

The stability problem shows several most interesting results with the introduction of the additional internal damping parameter for the standard linear model as opposed to the single internal damping parameter of the Kelvin-Voigt model. Several interesting cross-plots of flutter load versus both internal and external damping parameters show a considerable variation of flutter load with damping parameters and the ability to determine a maximum flutter load for the case of any one damping parameter fixed with the ability of allowing the other two damping parameters to vary until the maximum is achieved.

The basic optimization procedure was to fix or choose a desired flutter load and then numerically determine the optimum design which minimizes the beam mass. A special class of generalized parabola type boundary curves with beam thickness being a function of axial displacement was applied to Hauger's problem for a beam of rectangular cross-section. Important constraints considered are that the Euler Beam Theory is only valid for a thickness to length ratio smaller than 1/10 th and that practically the free end of the beam must be of finite thickness. Graphical results showed

that considerable weight reductions were possible and yielded excellent starting values for the optimization procedure. A Rosenbrock optimization routine with a minimum tip thickness constraint imposed was then applied to several beams with different values of internal and external damping parameters with mass reductions in the range of 20% to in excess of 30% resulting.

2. THE EQUATION OF MOTION. The equation of motion for the vibration of Euler beams with both internal and external damping was derived in reference [1]. The purpose of this section will be to develop similar equations which now include stability terms.

Following Brunelle [2], displacements are assumed of the form

$$\begin{aligned} v &= z\,\psi(y,t) \\ \bar{w} &= w_o(y) + w(y,t), \end{aligned} \tag{1}$$

the strain-displacement relations are

$$\begin{aligned} \varepsilon_y &= z\,\frac{\partial \psi}{\partial y} \\ \varepsilon_z &= 0 \\ \varepsilon_{yz} &= \frac{1}{2}\left[\frac{\partial w}{\partial y} + \psi\right], \end{aligned} \tag{2}$$

and the stress-strain law becomes

$$\sigma_y = E\,z\,\frac{\partial \psi}{\partial y} = -\,E\,z\,\frac{\partial^2 w}{\partial y^2} \tag{3}$$

if one sets $\varepsilon_{yz} = 0$ to allow for the equation

$$\psi = -\,\partial w/\partial y. \tag{4}$$

Thus, the averaged moment equation becomes

$$M = -\int_{A^*} z\,\sigma_y\,dA^* = E\,I\,\frac{\partial^2 w}{\partial y^2} \tag{5}$$

and consequently from (3) stress may be expressed as

$$\sigma_y = -\,\frac{Mz}{I}\,, \tag{6}$$

where I is the second moment of inertia about the neutral axis.

Taking a deformed beam element with sides perpendicular to the deflected centerline [3], Figure 1, a sum of vertical forces and moments yields the equations

$$(7) \qquad q - m\ddot{w} - F_{D1} - F_{D2} + \frac{\partial Q}{\partial y} + N_y \frac{\partial^2 \bar{w}}{\partial y^2} = 0$$

$$(8) \qquad Q = - \frac{\partial M}{\partial y}$$

where Q is the shear resultant, N_y is the in-plane load, q is a distributed beam load, m is the beam mass, and the F_{Di}, i = 1,2, are external dampings. From Baker, Woolam, and Yound [1] it is known that

$$(9) \qquad F_{D1} = \frac{C_1}{\ell} \left|\frac{\partial w}{\partial t}\right| \frac{\partial w}{\partial t}$$

$$(10) \qquad F_{D2} = \frac{C_2}{\ell} \frac{\partial w}{\partial t}$$

with the details of C_1 and C_2 being amply discussed in that reference.

Applying equations (9) and (10) to equation (7),

$$(11) \qquad q - m\ddot{w} - \frac{C_1}{\ell} \left|\frac{\partial w}{\partial t}\right| \frac{\partial w}{\partial t} - \frac{C_2}{\ell} \frac{\partial w}{\partial t} + \frac{\partial Q}{\partial y} + N_y \frac{\partial^2 \bar{w}}{\partial y^2} = 0.$$

Now, consider internal viscous damping according to the standard linear model

$$(12) \qquad (1 + C\, \partial/\partial t)\sigma_y = (E + E^*\, \partial/\partial t)\varepsilon_y$$

where E is Young's modulus and C and E* are viscoelastic material constants. Combining equations (2) and (4) results in

$$(13) \qquad \varepsilon_y = - z \frac{\partial^2 w}{\partial y^2}$$

and a direct substitution of this equation into equation (12) yields

$$(14) \qquad (1 + C \frac{\partial}{\partial t})\sigma_y = - z(E + E^* \frac{\partial}{\partial t}) \frac{\partial^2 w}{\partial y^2} .$$

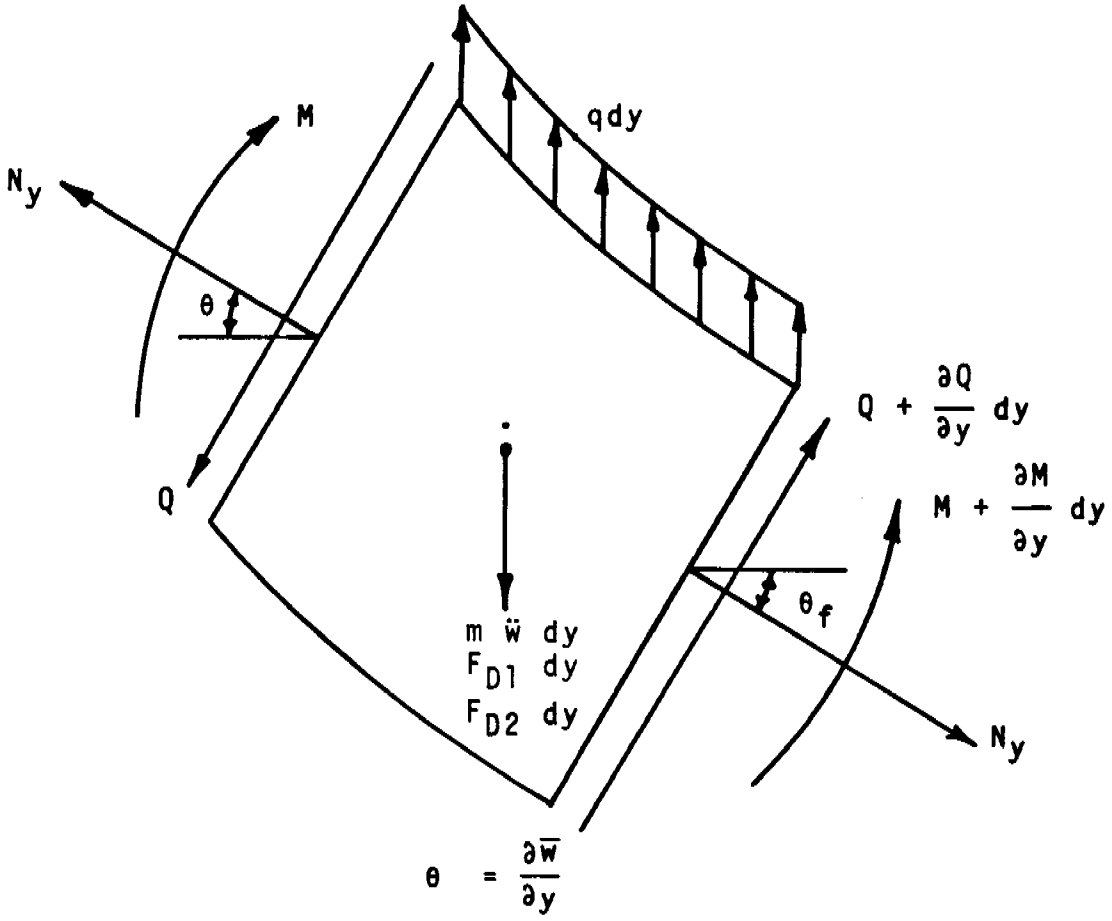


Figure 1. THE BEAM ELEMENT

Multiplying (14) by z and integrating over the cross-sectional beam area results in

$$(15) \qquad (1 + C \frac{\partial}{\partial t})M = (E + E^* \frac{\partial}{\partial t})I \frac{\partial^2 w}{\partial y^2} .$$

Now, assuming that C is independent of y, taking a partial derivative with respect to y of equation (15) results in

$$(16) \qquad (1 + C \frac{\partial}{\partial t}) \frac{\partial M}{\partial y} = \frac{\partial}{\partial y} [(E + E^* \frac{\partial}{\partial t})I \frac{\partial^2 w}{\partial y^2}]$$

and multiplying equation (8) by $(1 + C\ \partial()/\partial t)$ yields

$$(17) \qquad (1 + C \frac{\partial}{\partial t})Q = - (1 + C \frac{\partial}{\partial t}) \frac{\partial M}{\partial y} .$$

Hence,

$$(18) \qquad (1 + C \frac{\partial}{\partial t})Q = - \frac{\partial}{\partial y} [(E+E^* \frac{\partial}{\partial t})I \frac{\partial^2 w}{\partial y^2}$$

and a first derivative of this equation is

$$(19) \qquad (1 + C \frac{\partial}{\partial t}) \frac{\partial Q}{\partial y} = - \frac{\partial^2}{\partial y^2} [(E + E^* \frac{\partial}{\partial t})I \frac{\partial^2 w}{\partial y^2}].$$

Multiplying equation (11) by $(1 + C\ \partial()/\partial t)$

$$(1+C \frac{\partial}{\partial t})q - (1+C \frac{\partial}{\partial t})m\ \ddot{w} - (1+C \frac{\partial}{\partial t}) \frac{C_1}{\ell} |\frac{\partial w}{\partial t}| \frac{\partial w}{\partial t}$$

$$(20) \qquad - (1+C \frac{\partial}{\partial t}) \frac{C_2}{\ell} \frac{\partial w}{\partial t} + (1+C \frac{\partial}{\partial t}) \frac{\partial Q}{\partial y} + (1+C \frac{\partial}{\partial t})N_y \frac{\partial^2 \overline{w}}{\partial y^2} = 0$$

and a direct substitution of equation (19) into equation (20) results in the non-linear displacement equation of motion

$$\frac{\partial^2}{\partial y^2}\left[(E+E^*\frac{\partial}{\partial t})I\frac{\partial^2 w}{\partial y^2}\right] - (1+C\frac{\partial}{\partial t})N_y\frac{\partial^2 w}{\partial y^2}$$

$$(21)\qquad -(1+C\frac{\partial}{\partial t})q + (1+C\frac{\partial}{\partial t})\frac{C_1}{\ell}\left|\frac{\partial w}{\partial t}\right|\frac{\partial w}{\partial t}$$

$$+(1+C\frac{\partial}{\partial t})\frac{C_2}{\ell}\frac{\partial w}{\partial t} + (1+C\frac{\partial}{\partial t})m\ddot{w} = 0.$$

Equation (21) may be linearized by setting

$$(22)\qquad \left|\frac{\partial w}{\partial t}\right|\frac{\partial w}{\partial t} = 0$$

and hence the linear form of the equation is

$$\frac{\partial^2}{\partial y^2}\left[(E+E^*\frac{\partial}{\partial t})I\frac{\partial^2 w}{\partial y^2}\right] - (1+C\frac{\partial}{\partial t})N_y\frac{\partial^2 w}{\partial y^2}$$

$$(23)\qquad -(1+C\frac{\partial}{\partial t})q + (1+C\frac{\partial}{\partial t})\frac{C_2}{\ell}\frac{\partial w}{\partial t} + (1+C\frac{\partial}{\partial t})m\ddot{w} = 0.$$

In the case of elastic stability, a compressive load is applied and the substitution

$$(24)\qquad N_y = -P$$

is made. Also, since the material parameter C has been assumed independent of y it is now also assumed for consistency that the elastic and material parameters E and E* are also independent of y.

The corresponding boundary conditions for conservative flutter loads are

$$\left[E\frac{\partial}{\partial y}\left(I\frac{\partial^2 w}{\partial y^2}\right) + E^*\frac{\partial}{\partial y}\left(I\frac{\partial^3 w}{\partial y^2\partial t}\right) - N_y\frac{\partial w}{\partial y} -\right.$$

$$(25)\qquad \left. -C\,N_y\frac{\partial^2 w}{\partial y\partial t}\right] = 0 \qquad \text{or } w=0 \text{ on } x=0,\ell$$

and

$$\left[EI\,\frac{\partial^2 w}{\partial y^2} + E^*I\,\frac{\partial^3 w}{\partial y^2 \partial t}\right] = 0 \quad \text{or} \quad \frac{\partial w}{\partial y} = 0$$

(26) on $y=0,\ell$.

Special consideration is necessary for non-conservative flutter loads; for example, the boundary conditions for non-conservative flutter loads of cantilever Euler beams of length ℓ are for the clamped end

$$(27) \qquad w = 0, \quad \frac{\partial w}{\partial y} = 0 \quad \text{at } y = 0$$

and for the free end

$$(E+E^*\frac{\partial}{\partial t})\frac{\partial}{\partial y}(I\,\frac{\partial^2 w}{\partial y^2}) = 0,$$

$$(28) \qquad (E+E^*\frac{\partial}{\partial t})I\,\frac{\partial^2 w}{\partial y^2} = 0 \qquad \text{at } y = \ell.$$

3. THE ADJOINT VARIATIONAL PRINCIPLE. The adjoint to equation (23) will now be determined. Equation (23) in operator notation with $w_o = 0$ is

$$(29) \qquad L[w] = -q - C\,\frac{\partial q}{\partial t}$$

where

$$L[w] = -E\,\frac{\partial^2}{\partial y^2}\left[I\,\frac{\partial^2 w}{\partial y^2}\right] - E^*\,\frac{\partial^2}{\partial y^2}\left[I\,\frac{\partial^3 w}{\partial y^2 \partial t}\right]$$

$$+ N_y\,\frac{\partial^2 w}{\partial y^2} + C\,N_y\,\frac{\partial^3 w}{\partial y^2 \partial t} - \frac{C_2}{\ell}\,\frac{\partial w}{\partial t}$$

$$(30) \qquad -\left[\frac{C\,C_2}{\ell} + m\right]\frac{\partial^2 w}{\partial t^2} - C\,m\,\frac{\partial^3 w}{\partial t^3}\,.$$

Now, the adjoint equation

$$(31) \qquad L^*[v] = - q - C \frac{\partial q}{\partial t}$$

is sought.

In general, the Lagrange indentity

$$(32) \qquad v\, L[w] - wL^*[v] = \frac{d}{dy} P(w,v),$$

where $P(w,v)$ is the bilinear concomitant, can be integrated to yield the Green's formula

$$(33) \qquad \int_0^{\ell} (v\, L[w] - w\, L^*[v])dy = P(w,v)\Big|_0^{\ell}.$$

Notice from equations (29) and (31) that their right sides would cancel in applying equation (33), hence one may simply begin with the homogeneous form of operator (30),

$$(34) \qquad L[w] = 0.$$

A formal solution for the adjoint to operator (3) is obtained by starting with the equation

$$(35) \qquad \int_{t_0}^{t_1} \int_0^{\ell} v\, L[w]dy\, dt = 0$$

and manipulating by means of integration by parts towards the Green's formula (33). A direct substitution of operator (30) into equation (35) results in the equation

$$(36) \qquad \int_{t_0}^{t_1} \int_0^{\ell} v \left[\begin{array}{l} - E \frac{\partial^2}{\partial y^2} [I \frac{\partial^2 w}{\partial y^2}] - E^* \frac{\partial^2}{\partial y^2} [I \frac{\partial^3 w}{\partial y^2 \partial t}] \\ \\ + N_y \frac{\partial^2 w}{\partial y^2} + C\, N_y \frac{\partial^3 w}{\partial y^2 \partial t} - \frac{C_2}{\ell} \frac{\partial w}{\partial t} \\ \\ - [\frac{C\, C_2}{\ell} + m] \frac{\partial^2 w}{\partial t^2} - C\, m \frac{\partial^3 w}{\partial t^3} \end{array} \right] dy\, dt = 0;$$

omitting the details of numerous integrations by parts, the eventual result is

$$\int_{t_0}^{t_1}\int_0^{\ell}\left[v\left[\begin{array}{l} -E\,\frac{\partial}{\partial y^2}\left[I\,\frac{\partial^2 w}{\partial y^2}\right] - E^*\,\frac{\partial^2}{\partial y^2}\left[I\,\frac{\partial^3 w}{\partial y^2 \partial t}\right] \\ + N_y\,\frac{\partial^2 w}{\partial y^2} + C\,N_y\,\frac{\partial^3 w}{\partial y^2 \partial t} - \frac{C_2}{\ell}\,\frac{\partial w}{\partial t} \\ - \left[\frac{C\,C_2}{\ell} + m\right]\frac{\partial^2 w}{\partial t^2} - C\,m\,\frac{\partial^3 w}{\partial t^3} \end{array}\right] - w\left[\begin{array}{l} E\,\frac{\partial^2}{\partial y^2}\left[I\,\frac{\partial^2 v}{\partial y^2}\right] - E^*\,\frac{\partial^2}{\partial y^2}\left[I\,\frac{\partial^3 v}{\partial y^2 \partial t}\right] \\ - \frac{\partial^2}{\partial y^2}\left(N_y v\right) + C\,\frac{\partial^3}{\partial y^2 \partial t}\left(v\,N_y\right) \\ - \frac{C_2}{\ell}\,\frac{\partial v}{\partial t} + \left(\frac{C\,C_2}{\ell} + m\right)\frac{\partial^2 v}{\partial t^2} - C\,m\,\frac{\partial^3 v}{\partial t^3} \end{array}\right]\right] dy\,dt =$$

(37)

$$= \int_{t_0}^{t_1}\left[\begin{array}{l} \left[E\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^2 w}{\partial y^2}\right)v\right] - E\left(I\,\frac{\partial^2 w}{\partial y^2}\right)\frac{\partial v}{\partial y} \\ + E\left(I\,\frac{\partial^2 v}{\partial y^2}\right)\frac{\partial w}{\partial y} - E\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^2 v}{\partial y^2}\right)w \\ + E^*\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^3 w}{\partial y^2 \partial t}\right)v - E^*\left(I\,\frac{\partial^3 w}{\partial y^2 \partial t}\right)\frac{\partial v}{\partial y} \\ + E^*\left(I\,\frac{\partial^2 v}{\partial y^2}\right)\frac{\partial^2 w}{\partial y \partial t} - E^*\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^2 v}{\partial y^2}\right)\frac{\partial w}{\partial t} \\ - v\,N_y\,\frac{\partial w}{\partial t} + \frac{\partial}{\partial y}\left(N_y\,v\right)w \\ - C\,v\,N_y\,\frac{\partial^2 w}{\partial y \partial t} + C\,\frac{\partial}{\partial y}\left(v\,N_y\right)\frac{\partial w}{\partial t} \end{array}\right]_0^{\ell} dt$$

$$+\int_0^{\ell}\left[\begin{array}{l} E^* w \frac{\partial^2}{\partial y^2}\left(I \frac{\partial^3 v}{\partial y^2 \partial t}\right) - C \frac{\partial^2}{\partial y^2}(v N_y) w \\ + \frac{C_2}{\ell}(v w) + \left(\frac{C C_2}{\ell} + m\right) v \frac{\partial w}{\partial t} \\ - \left(\frac{C C_2}{\ell} + m\right) \frac{\partial v}{\partial t} w + C m v \frac{\partial^2 w}{\partial t^2} \\ - C m \frac{\partial v}{\partial t} \frac{\partial w}{\partial t} + C m \frac{\partial^2 v}{\partial t^2} w \end{array}\right]_{t_0}^{t_1} dy;$$

where the variables E^*, I, N_y, C, C_2, and m were assumed independent of time to arrive at this state.

Now, without loss of generality the usual restrictions that

$$w = \frac{\partial w}{\partial t} = v = 0 \qquad \text{at} \quad t = t_0, t_1 \tag{38}$$

are imposed on equation (37) with the result being that

$$\int_0^{\ell}\left[\begin{array}{l} E^* w \frac{\partial^2}{\partial y^2}\left(I \frac{\partial^3 v}{\partial y^2 \partial t}\right) - C \frac{\partial^2}{\partial y^2}(v N_y) w \\ + \frac{C_2}{\ell}(v w) + \left(\frac{C C_2}{\ell} + m\right) v \frac{\partial w}{\partial t} \\ - \left(\frac{C C_2}{\ell} + m\right) \frac{\partial v}{\partial t} w + C m v \frac{\partial^2 w}{\partial t^2} \\ - C m \frac{\partial v}{\partial t} \frac{\partial w}{\partial t} + C m \frac{\partial^2 v}{\partial t^2} w \end{array}\right]_{t_0}^{t_1} dy = 0. \tag{39}$$

Substituting (39) into equation (37) and comparing the result with Green's formula (33) results in the formula adjoint being

$$L^*[v] = E\frac{\partial^2}{\partial y^2}\left[I\frac{\partial^2 v}{\partial y^2}\right] - E^*\frac{\partial^2}{\partial y^2}\left[I\frac{\partial^3 v}{\partial y^2 \partial t}\right]$$

$$(40) \qquad -\frac{\partial^2}{\partial y^2}(N_y\, v) + C\frac{\partial^3}{\partial y^2 \partial t}(v\, N_y)$$

$$-\frac{C_2}{\ell}\frac{\partial v}{\partial t} + \left(\frac{C\,C_2}{\ell} + m\right)\frac{\partial^2 v}{\partial t^2} - C\,m\frac{\partial^3 v}{\partial t^3}$$

and the bilinear concomitant being

$$P(w,v)\Big|_0^{\ell} = \Bigg[E\frac{\partial}{\partial y}\left(I\frac{\partial^2 w}{\partial y^2}\right)v - E\,I\frac{\partial^2 w}{\partial y^2}\frac{\partial v}{\partial y}$$

$$+ E\,I\frac{\partial^2 v}{\partial y^2}\frac{\partial w}{\partial y} - E\frac{\partial}{\partial y}\left(I\frac{\partial^2 v}{\partial y^2}\right)w$$

$$(41) \qquad + E^*\frac{\partial}{\partial y}\left(I\frac{\partial^3 w}{\partial y^2 \partial t}\right)v - E^*I\frac{\partial^3 w}{\partial y^2 \partial t}\frac{\partial v}{\partial y}$$

$$+ E^*I\frac{\partial^2 v}{\partial y^2}\frac{\partial^2 w}{\partial y \partial t} - E^*\frac{\partial}{\partial y}\left(I\frac{\partial^2 w}{\partial y^2}\right)\frac{\partial w}{\partial t}$$

$$- v\,N_y\frac{\partial w}{\partial t} + \frac{\partial}{\partial y}(N_y\, v)w$$

$$- C\,v\,N_y\frac{\partial^2 w}{\partial y \partial t} + C\frac{\partial}{\partial y}(v\,N_y)\frac{\partial w}{\partial t} \Bigg]_0^{\ell}$$

The original and adjoint boundary conditions are now determined by setting the bilinear concomitant, equation (41), equal to zero. The problem here is in deciding which terms to group out as the original boundary conditions and consequently which terms will group out as the adjoint boundary conditions -- a simple inspection of equation (41) indicates that several choices are possible.

In the case of the non-conservative flutter loads of cantilever Euler beams, the original boundary conditions were postulated in the form of equations (27) and (28). Based upon equation (28), the expressions

$$\left[E \frac{\partial}{\partial y}\left(I \frac{\partial^2 w}{\partial y^2}\right) + E^* \frac{\partial}{\partial y}\left(I \frac{\partial^3 w}{\partial y^2 \partial t}\right)\right] v \Big|_0^{\ell} = 0$$

$$\left[E\, I \frac{\partial^2 w}{\partial y^2} + E^* I \frac{\partial^3 w}{\partial y^2 \partial t}\right] \frac{\partial v}{\partial y} \Big|_0^{\ell} = 0 \tag{42}$$

may be directly isolated from the bilinear concomitant (41) with a corresponding reduction to

$$P(w,v)\Big|_0^{\ell} = \left[\begin{array}{l} E\, I \dfrac{\partial^2 v}{\partial y^2} \dfrac{\partial w}{\partial y} - E \dfrac{\partial}{\partial y}\left(I \dfrac{\partial^2 v}{\partial y^2}\right) w \\ \\ + E^* I \dfrac{\partial^2 v}{\partial y^2} \dfrac{\partial^2 w}{\partial y \partial t} - E^* \dfrac{\partial}{\partial y}\left(I \dfrac{\partial^2 v}{\partial y^2}\right) \dfrac{\partial w}{\partial t} \\ \\ - v\, N_y \dfrac{\partial w}{\partial t} + \dfrac{\partial}{\partial y}\left(N_y\, v\right) w \\ \\ - C\, v\, N_y \dfrac{\partial^2 w}{\partial y \partial t} + C \dfrac{\partial}{\partial y}\left(v\, N_y\right) \dfrac{\partial w}{\partial t} \end{array}\right]_0^{\ell} \tag{43}$$

Now, equation (27) dictates that the remaining terms in (43) group out as products of either w or $\partial w/\partial y$; if this is to occur, the terms with time derivatives of w must be put into a different form. To accomplish this, equation (43) is integrated with respect to time to yield

$$\int_{t_0}^{t_1} P(w,v)\Big|_0^{\ell} dt = \int_{t_0}^{t_1} \left[\begin{array}{l} EI \dfrac{\partial^2 v}{\partial y^2} \dfrac{\partial w}{\partial y} - E \dfrac{\partial}{\partial y}\left(I \dfrac{\partial^2 v}{\partial y^2}\right) w \\ \\ + E^* I \dfrac{\partial^2 v}{\partial y^2} \dfrac{\partial^2 w}{\partial y \partial t} - E^* \dfrac{\partial}{\partial y}\left(I \dfrac{\partial^2 v}{\partial y^2}\right) \dfrac{\partial w}{\partial t} \\ \\ - v\, N_y \dfrac{\partial w}{\partial t} + \dfrac{\partial}{\partial y}\left(N_y\, v\right) w \\ \\ - C\, v\, N_y \dfrac{\partial^2 w}{\partial y \partial t} + C \dfrac{\partial}{\partial y}\left(v\, N_y\right) \dfrac{\partial w}{\partial t} \end{array}\right]_0^{\ell} dt \tag{44}$$

and those terms with time derivatives of w are integrated by parts with equation (38) appropriately applied to result in

$$(45)\qquad \int_{t_0}^{t_1} P(w,v)\Big|_0^{\ell}\, dt =$$

$$= \int_{t_0}^{t_1} \left[\left[EI\,\frac{\partial^2 v}{\partial y^2} - E^*I\,\frac{\partial^3 v}{\partial y^2 \partial t} - v\,N_y + C\,N_y\,\frac{\partial v}{\partial t} \right] \frac{\partial w}{\partial y} + \left[-E\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^2 v}{\partial y^2}\right) + E^*\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^3 v}{\partial y^2 \partial t}\right) + \frac{\partial}{\partial y}(N_y v) - C\,\frac{\partial^2}{\partial t \partial y}(v\,N_y) \right] w \right]_0^{\ell} dt.$$

From equation (45), it is clear that one can set

$$\left[EI\,\frac{\partial^2 v}{\partial y^2} - E^*I\,\frac{\partial^3 v}{\partial y^2 \partial t} - v\,N_y + C\,N_y\,\frac{\partial v}{\partial t} \right] \frac{\partial w}{\partial y}\Big|_0^{\ell} = 0$$

$$(46)\qquad \left[-E\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^2 v}{\partial y^2}\right) + E^*\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^3 v}{\partial y^2 \partial t}\right) + \frac{\partial}{\partial y}(N_y v) - C\,\frac{\partial^2}{\partial t \partial y}(v\,N_y) \right] w\Big|_0^{\ell} = 0$$

and as a result of this the bilinear concomitant is zero

$$(47)\qquad P(w,v)\Big|_0^{\ell} = 0$$

such that the formal adjoint (40) may now properly just be termed the adjoint. Since equations (27) and (28) are the original boundary conditions, their satisfaction in equations (42) and (46) leaves the equations

$$(48)\qquad v = 0, \quad \frac{\partial v}{\partial y} = 0 \qquad \text{on } y = 0$$

and

$$\left[EI\,\frac{\partial^2 v}{\partial y^2} - E^*I\,\frac{\partial^3 v}{\partial y^2 \partial t} - v\,N_y + C\,N_y\,\frac{\partial v}{\partial t} \right] = 0$$

$$\left[+E\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^2 v}{\partial y^2}\right) - E^*\,\frac{\partial}{\partial y}\left(I\,\frac{\partial^3 v}{\partial y^2 \partial t}\right) - \frac{\partial}{\partial y}(N_y v) + \right.$$

$$\left. + \frac{C\,\partial^2}{\partial t \partial y}(v\,N_y \right] = 0$$

$$(49)\qquad \text{on} \quad y = \ell$$

for the adjoint boundary conditions.

4. THE DIMENSIONLESS PROBLEM. At this point the original and adjoint differential equations together with their respective boundary conditions are known. Before proceeding on to the adjoint variational principle, it would be advantageous to develop a dimensionless formulation of the problem at hand and to discuss possible types of flutter load.

To begin with, the dimensionless variables

$$x = y/\ell$$
$$\tau = t/\sigma \tag{50}$$

are taken and it is assumed that the beam area and moment of inertia may take the form

$$I(x) = I_0 \tilde{I}(x)$$
$$A(x) = A_0 \tilde{A}(x). \tag{51}$$

Also, the compressive load $N_y = -P$, equation (24), is introduced into the equations at this point. For general convenience, the following dimensionless variables are applied throughout the development

$$\alpha(x) = \tilde{I}(x)$$
$$\gamma = E^*/E\sigma$$
$$p(x) = P(x)\ell^2/EI_0$$
$$\xi = C/\sigma$$
$$\bar{q} = C_2\ell^3/EI_0\sigma \tag{52}$$
$$\beta(x) = \rho A_0 \ell^4 \tilde{A}(x)/EI_0\sigma^2.$$

Thus, a direct application of equations (50-52) to equations (27), (28), (29), and (30) together with a bit of manipulation and the assumption that

$$q = 0 \tag{53}$$

results in the dimensionless original differential equation

$$\frac{\partial^2}{\partial x^2}\left[\alpha(x)\frac{\partial^2 w}{\partial x^2}\right] + \gamma\frac{\partial^2}{\partial x^2}\left[\alpha(x)\frac{\partial^3 w}{\partial x^2 \partial \tau}\right] + p(x)\frac{\partial^2 w}{\partial x^2}$$
$$+ \xi p(x)\frac{\partial^3 w}{\partial x^2 \partial \tau} + \bar{q}\frac{\partial w}{\partial \tau} + [\xi\bar{q}+\beta(x)]\frac{\partial^2 w}{\partial \tau^2} + \xi\beta(x)\frac{\partial^3 w}{\partial \tau^3} = 0 \tag{54}$$

together with the non-conservative dimensionless original boundary conditions

$$w = 0, \quad \frac{\partial w}{\partial x} = 0 \qquad \text{on } x=0 \tag{55}$$

and

$$[1 + \gamma \frac{\partial}{\partial \tau}]\alpha(x) \frac{\partial^2 w}{\partial x^2} = 0$$

$$[1 + \gamma \frac{\partial}{\partial \tau}] \frac{\partial}{\partial x} [\alpha(x) \frac{\partial^2 w}{\partial x^2}] = 0 \qquad \text{on } x=1 \tag{56}$$

for a cantilevered Euler beam. Similarly, equations (50-53) applied to equations (31), (40), (48), and (49) result in the dimensionless adjoint differential equation

$$\frac{\partial^2}{\partial x^2} [\alpha(x) \frac{\partial^2 v}{\partial x^2}] - \gamma \frac{\partial^2}{\partial x^2} [\alpha(x) \frac{\partial^3 v}{\partial x^2 \partial \tau}] + \frac{\partial^2}{\partial x^2} [p(x)v]$$

$$- \xi \frac{\partial^3}{\partial x^2 \partial \tau} [p(x)v] - \bar{q} \frac{\partial v}{\partial \tau} + [\xi \bar{q} + \beta(x)] \frac{\partial^2 v}{\partial \tau^2} - \xi \beta(x) \frac{\partial^3 v}{\partial \tau^3} = 0 \tag{57}$$

and the dimensionless adjoint boundary conditions

$$v = 0, \quad \frac{\partial v}{\partial x} = 0 \qquad \text{on } x=0 \tag{58}$$

and

$$\alpha(x) \frac{\partial^2 v}{\partial x^2} - \gamma \alpha(x) \frac{\partial^3 v}{\partial x^2 \partial \tau} + p(x)v - \xi p(x) \frac{\partial v}{\partial \tau} = 0$$

$$\frac{\partial}{\partial x} [\alpha(x) \frac{\partial^2 v}{\partial x^2}] - \gamma \frac{\partial}{\partial x} [\alpha(x) \frac{\partial^3 v}{\partial x^2 \partial \tau}] + \frac{\partial}{\partial x} [p(x)v]$$

$$- \xi \frac{\partial^2}{\partial x \partial \tau} [p(x)v] = 0 \qquad \text{on } x=1. \tag{59}$$

Some examples of the loading P(x) for several distinct flutter problems are

$$P(x) = P \tag{60}$$

for Beck's problem [4] which assumes a concentrated load is applied tangent to the free end,

$$P(x) = q(1-x) \tag{61}$$

for the Leipholz problem [5] which assumes that a uniformly distributed load acts tangentially along the centerline of the beam, and

$$P(x) = \frac{1}{2} q_0 (1-x)^2 \tag{62}$$

for Hauger's problem [6] which assumes that a linear distributed tangential load acts along the centerline of the beam. These are merely examples of what can be handled, the analysis can be applied to any problem which satisfies equation (54), with the P(x) now being specified for this problem, and the corresponding boundary conditions (55) and (56).

5. THE ADJOINT VARIATIONAL FORMULATION. Based upon a suggestion by Dr. Gary Anderson, considerable simplification in developing the variational principle is possible if solutions of the form

$$\begin{aligned} w(x,\tau) &= W(x)e^{\lambda\tau} \\ v(x,\tau) &= V(x)e^{-\lambda\tau} \end{aligned} \tag{63}$$

are immediately assumed. A direct application of solution (63) to the original problem (54-56) results in the differential equation

$$\begin{aligned} &\frac{\partial^2}{\partial x^2}\left[\alpha(x)[1+\lambda\gamma]\frac{\partial^2 W}{\partial x^2}\right] + [1+\lambda\xi]p(x)\frac{\partial^2 W}{\partial x^2} \\ &+ [\lambda\bar{q}(1+\lambda\xi) + \lambda^2\beta(x)[1+\lambda\xi]]W = 0 \end{aligned} \tag{64}$$

and the boundary conditions

$$W = 0, \quad \frac{\partial W}{\partial x} = 0, \quad \text{on} \quad x=0 \tag{65}$$

and

$$(1+\gamma\lambda)\alpha(x)\frac{\partial^2 W}{\partial x^2} = 0$$

(66)

$$(1+\gamma\lambda)\frac{\partial}{\partial x}\left[\alpha(x)\frac{\partial^2 W}{\partial x^2}\right] = 0 \qquad \text{on} \quad x=1.$$

Likewise, applying solution (63) to the adjoint problem (57-59) results in the differential equation

$$\frac{\partial^2}{\partial x^2}\left[(1+\gamma\lambda)\alpha(x)\frac{\partial^2 V}{\partial x^2}\right] + \frac{\partial^2}{\partial x^2}\left[(1+\lambda\xi)p(x)V\right]$$

(67) $$+ \left[\lambda\bar{q}(1+\lambda\xi) + \lambda^2\beta(x)\left[1+\xi\lambda\right]\right]V = 0$$

and the boundary conditions

(68) $$V = 0, \qquad \frac{\partial V}{\partial x} = 0 \qquad \text{on} \quad x=0$$

and

$$(1+\gamma\lambda)\alpha(x)\frac{\partial^2 V}{\partial x^2} + (1+\lambda\xi)p(x)V = 0$$

(69) $$\frac{\partial}{\partial x}\left[(1+\lambda\gamma)\alpha(x)\frac{\partial^2 V}{\partial x^2}\right] + \frac{\partial}{\partial x}\left[(1+\lambda\xi)p(x)V\right] = 0 \qquad \text{on} \quad x=1.$$

Now, a variational principle based upon equations (64-69) will be developed in terms of potential energy V^* such that the ultimate result is

(70) $$\delta V^* = 0.$$

To begin with, equation (64) is multiplied by δV and integrated over x

$$\int_0^1 \Bigg[\frac{\partial^2}{\partial x^2}\left[(1+\lambda\gamma)\alpha(x)\frac{\partial^2 W}{\partial x^2}\right]\delta V$$

(71) $$+ \left[1+\lambda\xi\right]p(x)\frac{\partial^2 W}{\partial x^2}\delta V$$

$$+ \left[(1+\lambda\xi)\lambda\bar{q} + (1+\lambda\xi)\lambda^2\beta(x)\right]W\delta V \Bigg] \, dx=0.$$

After an application of integration by parts and a bit of manipulation, equation (71) may be put into the form

$$\delta \int_0^1 \left[[1+\lambda\gamma)\alpha(x) \frac{\partial^2 W}{\partial x^2} \frac{\partial^2 V}{\partial x^2} - [p(x) \frac{\partial W}{\partial x} \frac{\partial V}{\partial x} + \frac{\partial p(x)}{\partial x} \frac{\partial W}{\partial x} V] (1+\lambda\xi) + [(1+\lambda\xi)\lambda\bar{q} + (1+\lambda\xi)\lambda^2\beta(x)]WV \right] dx$$

$$+ \left[\left(\frac{\partial}{\partial x} [(1+\lambda\xi)\alpha(x) \frac{\partial^2 W}{\partial x^2}] + (1+\lambda\xi) \frac{\partial W}{\partial x} p(x) \right) \delta V + \left(\frac{\partial}{\partial x} [(1+\lambda\gamma)\alpha(x) \frac{\partial^2 V}{\partial x^2}] + \frac{\partial}{\partial x} [(1+\lambda\xi)p(x)V] \right) \delta W - [(1+\lambda\gamma)\alpha(x) \frac{\partial^2 W}{\partial x^2}] \delta(\frac{\partial W}{\partial x}) - [(1+\lambda\gamma)\alpha(x) \frac{\partial^2 V}{\partial x^2}] \delta(\frac{\partial V}{\partial x}) \right]_0^1 = 0. \quad (72)$$

In looking at equation (72), it is clear from equations (65-66) and (68-69) that the following boundary conditions are immediately satisfied

$$V = W = \frac{\partial V}{\partial x} = \frac{\partial W}{\partial x} = 0 \qquad \text{on} \quad x=0$$

$$(1+\lambda\gamma)\alpha(x) \frac{\partial^2 W}{\partial x^2} = 0 \qquad \text{on} \quad x=1 \quad (73)$$

$$\frac{\partial}{\partial x} [(1+\lambda\gamma)\alpha(x) \frac{\partial^2 V}{\partial x^2}] + \frac{\partial}{\partial x} [(1+\lambda\xi)p(x)V] = 0 \qquad \text{on} \quad x=1$$

$$(1+\lambda\gamma) \frac{\partial}{\partial x} [\alpha(x) \frac{\partial^2 W}{\partial x^2}] = 0 \qquad \text{on} \quad x=1$$

and that the boundary condition

$$(74) \qquad (1+\gamma\lambda)\alpha(x)\frac{\partial^2 V}{\partial x^2} + (1+\lambda\xi)p(x)V = 0 \qquad \text{on } x=1$$

must still be satisfied. If it is now stipulated that boundary condition (74) must also be satisfied, equation (72) can be put into the form

$$\delta \int_0^1 \left[\begin{array}{l} [(1+\lambda\gamma)\alpha(x)\dfrac{\partial^2 W}{\partial x^2}\dfrac{\partial^2 V}{\partial x^2}] \\ \\ -(1+\lambda\xi)\dfrac{\partial W}{\partial x}\dfrac{\partial}{\partial x}[p(x)V] \\ \\ +[(1+\lambda\xi)\lambda\overline{q} + (1+\lambda\xi)\lambda^2\beta(x)]WV \end{array} \right] dx$$

$$(75) \qquad + \delta[(1+\lambda\xi)p(x)\frac{\partial W}{\partial x}V]_{x=1} = 0.$$

To arrive at the variational principle (70), it is now assumed that the potential energy may be broken up as

$$(76) \qquad V^* = V_1^* + V_2^*,$$

where

$$(77) \qquad V_1^* = \frac{1}{2}\int_0^1 V_1\, dx$$

$$(78) \qquad V_2^* = \int_0^1 V_2\, \delta(x-1)dx$$

and equation (78) reflects the introduction of the Dirac Delta function $\delta(x-a)$ with

$$(79) \qquad \int_0^a \phi(x)\, \delta(x-a)dx = \phi(a)/2$$

being one of its most useful integral properties. Thus, applying equations (76-78) to a comparison of equations (70) and (75) results in the potentials

$$V_1 = [(1+\lambda\gamma)\alpha(x) \frac{\partial^2 W}{\partial x^2} \frac{\partial^2 V}{\partial x^2}]$$

$$- (1+\lambda\xi) \frac{\partial W}{\partial x} \frac{\partial}{\partial x} [p(x)V]$$

$$+ [(1+\lambda\xi)\lambda\bar{q} + (1+\lambda\xi)\lambda^2\beta(x)]WV \tag{80}$$

$$V_2 = (1+\lambda\xi)p(x) \frac{\partial W}{\partial x} V. \tag{81}$$

The calculus of variations will now be used to determine the Euler-Lagrange equations associated with equation (70). Clearly, the functional dependencies of V_1 and V_2 are

$$V_1 = V_1 \left(x : V, W, \frac{\partial V}{\partial x}, \frac{\partial W}{\partial x}, \frac{\partial^2 V}{\partial x^2}, \frac{\partial^2 W}{\partial x^2}\right)$$

$$V_2 = V_2 \left(x : V, \partial W/\partial x\right) \tag{82}$$

and it may be postulated that

$$I(V,W) = \int_0^1 V_1 \left(x : V, W, \frac{\partial V}{\partial x}, \frac{\partial W}{\partial x}, \frac{\partial^2 V}{\partial x^2}, \frac{\partial^2 W}{\partial x^2}\right) dx$$

$$+ \int_0^1 V_2 \left(x : V, \frac{\partial W}{\partial x}\right) \delta(x-1)dx \tag{83}$$

such that

$$\delta I = 0. \tag{84}$$

A back substitution of equation (84) into equation (83), integration by parts, and an application of the first lemma of the calculus of variations leads to the Euler-Lagrange equations

$$\begin{aligned} &\frac{\partial^2}{\partial x^2}\left(\frac{\partial V_1}{\partial V_{xx}}\right) - \frac{\partial}{\partial x}\left(\frac{\partial V_1}{\partial V_x}\right) + \frac{\partial V_1}{\partial V} = 0 \\ &\frac{\partial^2}{\partial x^2}\left(\frac{\partial V_1}{\partial W_{xx}}\right) - \frac{\partial}{\partial x}\left(\frac{\partial V_1}{\partial W_x}\right) + \frac{\partial V_1}{\partial W} = 0 \end{aligned} \tag{85}$$

and the boundary conditions

$$
\begin{aligned}
&-\frac{\partial}{\partial x}\left(\frac{\partial V_1}{\partial V_{xx}}\right)+\frac{\partial V_1}{\partial V_x}=0 && \text{or } V=0 \text{ on } x=0\\
&-\frac{\partial}{\partial x}\left(\frac{\partial V_1}{\partial V_{xx}}\right)+\frac{\partial V_1}{\partial V_x}+\frac{\partial V_2}{\partial V}=0 && \text{or } V=0 \text{ on } x=1\\
&\frac{\partial V_1}{\partial V_{xx}}=0 && \text{or } V_x=0 \text{ on } x=0,1\\
&-\frac{\partial}{\partial x}\left(\frac{\partial V_1}{\partial W_{xx}}\right)+\frac{\partial V_1}{\partial W_x}=0 && \text{or } W=0 \text{ on } x=0,1\\
&\frac{\partial V_1}{\partial W_{xx}}=0 && \text{or } W_x=0 \text{ on } x=0\\
&\frac{\partial V_1}{\partial W_{xx}}+\frac{\partial V_1}{\partial W_x}=0 && \text{or } W_x=0 \text{ on } x=1.
\end{aligned}
\tag{86}
$$

A direct substitution of equations (80-81) into equations (85-86) results in an exact reformulation of the original problem (64-66) as well as the adjoint problem (67-69). This verifies that the variational principle (70) is indeed correct when V^* is defined by equations (76-78) and (80-81).

6. THE GENERALIZED RITZ APPROXIMATION. Following Anderson [7], a natural extension of the Ritz method in its classical form is applied to the current adjoint boundary value problem. The approximation is made in terms of a finite series expansion of the product of prespecified coordinate functions. The approximation will now be made as the two series

$$
\begin{aligned}
W(x) &\simeq \sum_{n=1}^{N} a_n W_n(x)\\
V(x) &\simeq \sum_{n=1}^{N} b_n V_n(x)
\end{aligned}
\tag{87}
$$

where $W_n(x)$ and $V_n(x)$ are the coordinate functions which must be prescribed and a_n and b_n are constants. Thus, the general procedure involved will be to select $W_n(x)$ and $V_n(x)$ subject to certain boundary constraints and then to use the variational principle to find a characteristic equation for flutter load. In

general, it is believed that satisfaction of all or most of the boundary constraints tends to minimize convergence time [7]. However, if those terms in the natural boundary conditions which contain derivatives of odd order of time [7] are deleted, then the coordinate functions which were previously developed for the undamped beam may be employed directly in the current problem. Since the geometric boundary conditions remain unchanged, no extra consideration is warranted in this case. Thus, going back to natural boundary conditions (56) and (59) with the assumption that odd order derivatives in time are deleted and tracing through ensuing developments results in the boundary conditions

$$\alpha(x)\frac{\partial^2 w}{\partial x^2} = 0$$

$$\frac{\partial}{\partial x}\left[\alpha(x)\frac{\partial^2 w}{\partial x^2}\right] = 0 \qquad \text{on } x=1 \tag{88}$$

and

$$\alpha(x)\frac{\partial^2 v}{\partial x^2} + p(x)v = 0$$

$$\frac{\partial}{\partial x}\left[\alpha(x)\frac{\partial^2 v}{\partial x^2}\right] + \frac{\partial}{\partial x}\left[p(x)v\right] = 0 \qquad \text{on } x=1 \tag{89}$$

which are now utilized in place of boundary conditions (66) and (69).

Following reference [8], it is the current intention to choose the $W_n(y)$ and $V_n(y)$ as general polynomials. The polynomials will be chosen such that there is one term for each boundary condition plus one, the highest power is one more than the highest order derivative, and the n the order polynomial is of the form

$$x^{(n-1)}\,(a_1 + \cdots + a_5\,x^4). \tag{90}$$

Hence, the assumed forms of the coordinate functions are

$$W_n(x) = x^{(n-1)}(a_1 + a_2x + a_3x^2 + a_4x^3 + a_5x^4) \tag{91}$$

$$V_n(x) = x^{(n-1)}(b_1 + b_2x + b_3x^2 + b_4x^3 + b_5x^4). \tag{92}$$

Now, the constants in coordinate functions (91) and (92) are evaluated by substitution of these equations into boundary constraints (65), (88) and (68), (89), respectively, which they must identically satisfy. Thus, a direct substitution of equation (91) into the boundary conditions results in the four constants

$$a_1 = 0, \qquad a_2 = 0$$

$$a_3 = [(n+3)(n+2)^2(n+1)\alpha(1)]a_5/d \tag{93}$$

$$a_4 = [2(n+3)(n+2)(n+1)n\alpha(1)]a_5/d$$

where the denominator term d is defined as

$$d = (n+2)(n+1)^2 n\alpha(1). \tag{94}$$

Since the a_5 in equations (93) is really just an arbitrary constant if traced through to its origin in equation (91), substantial simplification is realized by choosing it as

$$a_5 = d/[(n+2)(n+3)\alpha(1)]. \tag{95}$$

A direct back-substitution of equations (93-95) into equation (91) results in the final form of the coordinate function being

$$W_n(x) = x^{(n+1)}[(n+3)(n+2) - 2(n+3)n\,x + (n+1)nx^2]. \tag{96}$$

Similarly, a direct substitution of equation (92) into the boundary conditions results in the four constants

$$b_1 = 0, \qquad b_2 = 0$$

$$b_3 = \begin{bmatrix} (n+3)(n+2)(n+1)\alpha^2(1) + 2n(n+2)\alpha(1)p(1) \\ - 2(n+2)\alpha(1)p'(1) + 2(n+2)\alpha'(1)p(1) + p^2(1) \end{bmatrix} b_5/\tilde{d} \tag{97}$$

$$b_4 = \begin{bmatrix} -2(n+3)(n+2)(n+1)n\,\alpha^2(1) - 4n(n+1)\alpha(1)p(1) \\ +2(2n+3)\alpha(1)p'(1) - 2(2n+3)\alpha'(1)p(1) - 2p^2(1) \end{bmatrix} b_5/\tilde{d}$$

where the denominator term $\tilde{d}$ is defined as

$$\tilde{d} = (n+2)(n+1)^2 n\,\alpha^2(1) + 2(n+1)(n-1)\alpha(1)p(1)$$

$$- 2(n+1)\alpha(1)p'(1) + 2(n+1)\alpha'(1)p(1) + p^2(1) \tag{98}$$

and the prime notation indicates derivatives with respect to coordinate x, i.e. $' = d/dx$.

Now, equations (97) reduce equation (92) to

$$V_n(x) = x^{(n+1)}[b_3 + b_4 x + b_5 x^2] \tag{99}$$

which upon letting $b_5 = \gamma$ results in

$$V_n(x) = x^{(n+1)}[\alpha_n - 2\beta_n x + \gamma_n x^2] \tag{100}$$

where

$$\begin{aligned}\alpha_n = {} & (n+3)(n+2)^2(n+1)\alpha^2(1) + 2n(n+2)\alpha(1)p(1) - 2(n+2)\alpha(1)p'(1) \\ & + 2(n+2)\alpha'(1)p(1) + p^2(1)\end{aligned} \tag{101}$$

$$\begin{aligned}\beta_n = {} & (n+3)(n+2)(n+1)n\alpha^2(1) + 2n(n+1)\alpha(1)p(1) - (2n+3)\alpha(1)p'(1) \\ & + (2n+3)\alpha'(1)p(1) + p^2(1)\end{aligned} \tag{102}$$

$$\begin{aligned}\gamma_n = {} & (n+2)(n+1)^2 n\,\alpha^2(1) + 2(n+1)(n-1)\alpha(1)p(1) - 2(n+1)\alpha(1)p'(1) \\ & + 2(n+1)\alpha'(1)p(1) + p^2(1).\end{aligned} \tag{103}$$

With the coordinate functions $W_n(x)$ and $V_n(x)$ now determined, it remains to solve for the flutter load $p(x)$. The approximate solutions (87) are now back substituted into the potential energy expression given by equations (76-81) to yield

$$\mathbf{V}^* = \frac{1}{2}\left[\begin{aligned}
& (1+\lambda\gamma)\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}\int_0^1 \alpha(x)W_k''(x)V_n''(x)dx \\
& -(1+\lambda\xi)\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}\int_0^1 p(x)W_k'(x)V_n'(x)dx \\
& -(1+\lambda\xi)\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}\int_0^1 p'(x)W_k'(x)V_n(x)dx \\
& +(1+\lambda\xi)\lambda\bar{q}\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}\int_0^1 W_k(x)V_n(x)dx \\
& +(1+\lambda\xi)\lambda^2\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}\int_0^1 \beta(x)W_k(x)V_n(x)dx
\end{aligned}\right]$$

$$+\frac{1}{2}(1+\lambda\xi)\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}p(x)W_k'(x)V_n(x)\Big|_{x=1} \tag{104}$$

where

$$G_{nk} = a_k\, b_n. \tag{105}$$

Now, a consideration of equation (104) shows that the following definitions are possible

$$A_{nk} = \int_0^1 \beta(x)W_k(x)V_n(x)dx \tag{106}$$

$$B_{nk} = \int_0^1 \alpha(x)W_k''(x)V_n''(x)dx \tag{107}$$

$$C_{nk} = \int_0^1 p(x)W_k'(x)V_n'(x)dx \tag{108}$$

$$D_{nk} = \int_0^1 p'(x)W_k'(x)V_n(x)dx \tag{109}$$

$$E_{nk} = \int_0^1 W_k(x)V_n(x)dx \tag{110}$$

$$F_{nk} = p(x)W_k'(x)V_n(x)\Big|_{x=1} \tag{111}$$

and hence equation (104) may be rewritten as

$$V^* = \frac{1}{2}\left[\begin{array}{l}(1+\lambda\gamma)\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}B_{nk}-(1+\lambda\xi)\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}C_{nk} \\ -(1+\lambda\xi)\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}D_{nk}+(1+\lambda\xi)\lambda\bar{q}\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}E_{nk} \\ +(1+\lambda\xi)\lambda^2\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}A_{nk}\end{array}\right]$$

$$+\frac{1}{2}(1+\lambda\xi)\sum_{k=1}^{k=K}\sum_{n=1}^{n=N} G_{nk}F_{nk}. \tag{112}$$

Thus, the variational principle (70) is now applied to equation (112) with the functional dependence of V^* being

$$V^* = V^*(a_k, b_n) \tag{113}$$

and the resulting Euler-Lagrange equations being

$$\frac{\partial V^*}{\partial a_k} = 0, \qquad \frac{\partial V^*}{\partial b_n} = 0. \tag{114}$$

Applying equations (114) to equation (112) and recalling from equation (105) that $G_{nk} = a_k b_n$ results in the conditions

$$\sum_{n=1}^{n=N}\begin{bmatrix} (1+\lambda\gamma)B_{nk} - (1+\lambda\xi)C_{nk} \\ -(1+\lambda\xi)D_{nk} + (1+\lambda\xi)\lambda\bar{q}\,E_{nk} \\ +(1+\lambda\xi)\lambda^2 A_{nk} + (1+\lambda\xi)F_{nk} \end{bmatrix} b_n = 0 \tag{115}$$

$$\sum_{k=1}^{k=K}\begin{bmatrix} (1+\lambda\gamma)B_{nk} - (1+\lambda\xi)\,C_{nk} \\ -(1+\lambda\xi)D_{nk} + (1+\lambda\xi)\lambda\bar{q}\,E_{nk} \\ +(1+\lambda\xi)\lambda^2\,A_{nk} + (1+\lambda\xi)\,F_{nk} \end{bmatrix} a_k = 0. \tag{116}$$

Clearly, equations (115) and (116) are equivalent and hence it is permissible from this point on to consider only one of these equations. Hence, working only with equation (115) it is possible to obtain the critical flutter load by setting the determinant of the K equations thus represented equal to zero with the resulting characteristic equation being

$$\mathrm{DET}\begin{vmatrix} (\xi\,A_{nk})\lambda^3 + (A_{nk} + \xi\bar{q}\,E_{nk})\lambda^2 \\ +(\gamma\,B_{nk} - \xi\,C_{nk} - \xi\,D_{nk} + \bar{q}\,E_{nk} + \xi\,F_{nk})\lambda \\ +(B_{nk} - C_{nk} - D_{nk} + F_{nk}) \end{vmatrix} = 0. \tag{117}$$

Now, equation (117) may be solved numerically on the digital computer for critical flutter load once the necessary material and geometric properties are specified.

7. THE OPTIMIZATION OF HAUGER'S PROBLEM. Hauger's problem [6] will now be optimized within the bounds of a special class of generalized parabola type boundary curves. The beam thickness will be expressed functionally in terms of length axis displacement, but the beam width b will be held constant and thus any cut perpendicular to the length axis will have a rectangular cross-section.

A generalized parabola type curve, Figure 2, is passed through the point z=a at the clamped end y=0 and the point z=c at the free end y=ℓ. The generalized form of this curve is chosen as

$$z - c = (\text{const.})(\ell - y)^{n} \tag{118}$$

and the constant is now evaluated such that at y=0, z=a with the result being

$$\text{const.} = \frac{(a-c)}{\ell^{n}} . \tag{119}$$

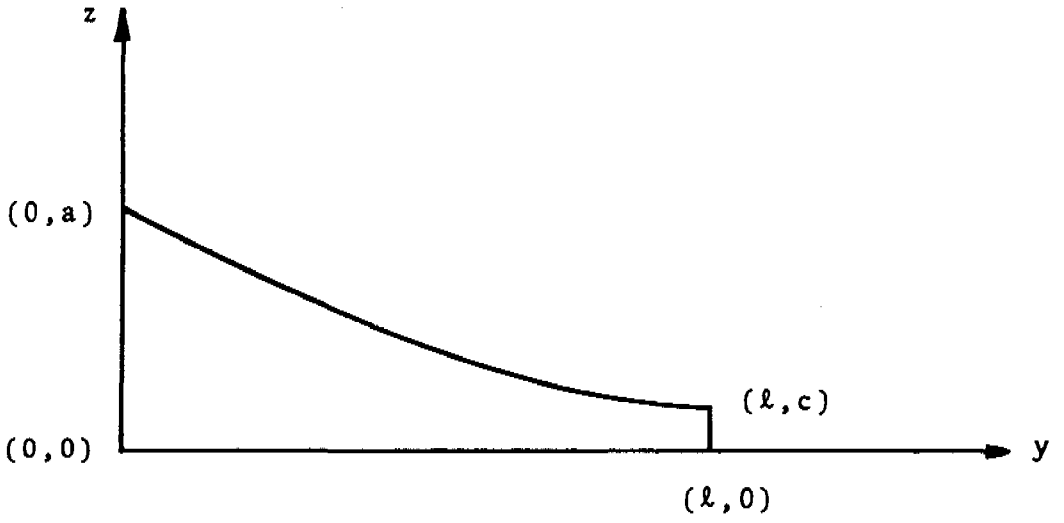


Figure 2. The Generalized Parabolic Type Boundary Curve

Representing the beam half-thickness as h(y) such that h(y)=z, then from (118) and (119)

$$(120) \qquad h(y) = c + (a-c)\left[\frac{\ell-y}{\ell}\right]^{n} .$$

The basic optimization problem to be considered is that of choosing or fixing the desired flutter load, in this case q_0, and then determining the optimum design h(y) which minimizes the beam mass.

The dimensionless variables (50) and (52) are introduced and following equations (51) the second moment of inertia is expressed as

$$(121) \qquad I(x) = I_0 S(x)$$

and in a similar vain the area can be written as

$$(122) \qquad A(x) = A_0 r(x).$$

Now, a suitable reference dimension will be introduced by using length as a characteristic beam reference and the dimensionless beam coordinates

$$(123) \qquad \alpha = c/\ell, \qquad \mu = a/\ell$$

are introduced and applied to equation (120) with the result being

$$(124) \qquad h(x) = \ell[\alpha + (\mu-\alpha)(1-x)^{n}].$$

The basic definition of second moment of inertia is evaluated for a constant beam width b and thickness 2h(x) to yield

$$(125) \qquad I = \int_A z^2 \, dA = \frac{2}{3} b \, h^3(x)$$

which upon an introduction of equation (124) becomes

$$(126) \qquad I = \frac{2}{3} b \, \ell^3 [\alpha + (\mu-\alpha)(1-x)^{n}]^3;$$

letting

$$(127) \qquad I_0 = \frac{2}{3} b \, \ell^3$$

and comparing equation (126) with equation (121) results in the definition

$$(128) \qquad S(x) = [\alpha + (\mu-\alpha)(1-x)^{n}]^3.$$

A similar look at the cross-sectional area

$$A(x) = 2b\,h(x) \tag{129}$$

with the definition

$$A_o = 2\,b\,\ell \tag{130}$$

being made results in equation (122) yielding the definition

$$r(x) = [\alpha + (\mu-\alpha)(1-x)^{\eta}]. \tag{131}$$

A comparison of equations (128) and (131) shows them to be functionally similar and if the function

$$f(x) = [\alpha + (\mu-\alpha)(1-x)^{\eta}] \tag{132}$$

is now defined, then

$$r(x) = f(x) \tag{133}$$

$$S(x) = f^3(x). \tag{134}$$

The basic optimization problem will be to minimize the beam mass while holding the flutter load constant, which can be expressed as

$$\int_0^1 \rho A(x)dx = \text{minimum}. \tag{135}$$

Equation (135) is readily expressed as

$$\int_0^1 \rho A(x)dx = 2\rho\ell^2 b \int_0^1 f(x)dx \tag{136}$$

and since ρ, ℓ, and b are essentially fixed variables the problem reduces to minimizing the integral

$$\int_0^1 f(x)dx = \text{minimum} \tag{137}$$

or hence minimizing

$$\int_0^1 f(x)dx = \alpha + [(\mu-\alpha)/(\eta+1)]. \tag{138}$$

8. THE APPROXIMATE SOLUTION. A first approximation to the current problem will be made by considering only the first several terms of the generalized Ritz solution for the original and the adjoint problems.

From equations (50), (52), and (62) it is clear that

$$p(x) = Q(1-x)^2 \tag{139}$$

where

$$Q = \frac{q_o \ell^4}{2 E I_o} \tag{140}$$

Also, from equation (50) and (52) and equations (51), (121), (122), (131), (133), and (134) it is clear that

$$\begin{aligned} \alpha(x) &= S(x) = f^3(x) \\ \beta(x) &= \hat{\beta}\, r(x) = \hat{\beta}\, f(x) \end{aligned} \tag{141}$$

where

$$\hat{\beta} = \frac{\rho A_o \ell^4}{E I_o \sigma^2} \tag{142}$$

and the now most convenient definition

$$\sigma^2 = \frac{\rho A_o \ell^4}{E I_o} \tag{143}$$

results in

$$\hat{\beta} = 1. \tag{144}$$

From equation (139) it is clear that

$$p(1) = p'(1) = 0 \tag{145}$$

and hence the constants (101-103) in the general Ritz coordinate functions (96) and (100) reduce to

$$\begin{aligned} \alpha_n &= (n+3)(n+2)^2(n+1)\alpha^2(1) \\ \beta_n &= (n+3)(n+2)(n+1)n\,\alpha^2(1) \\ \gamma_n &= (n+2)(n+1)^2 n\,\alpha^2(1). \end{aligned} \tag{146}$$

For later convenience in numerical calculations it is desirable that integrals (106)-(111) be independent of both α and μ and that the exact loading (139) now be introduced, hence equations (106)-(111) become

$$A_{nk} = \alpha\int_0^1 \bar{\gamma}W_k(x)V_n(x)dx + \mu\int_0^1 \bar{\xi}W_k(x)V_n(x)dx$$

$$B_{nk} = \alpha^3 \int_0^1 \bar{\gamma}^3 W_k'''(x)V_n'''(x)dx + 3\alpha^2\mu\int_0^1 \bar{\gamma}^2\bar{\xi}\, W_k'''(x)V_n'''(x)dx$$

$$+ 3\alpha\mu^2\int_0^1 \bar{\xi}^2\bar{\gamma}W_k'''(x)V_n'''(x)dx + \mu^3\int_0^1 \bar{\xi}^3 W_k'''(x)V_n'''(x)dx$$

$$C_{nk} = Q\int_0^1 (1-x)^2\, W_k'(x)\, V_n'(x)dx$$

$$D_{nk} = -2\,Q\int_0^1 (1-x)\, V_n(x)\, W_k'(x)dx$$

$$E_{nk} = \int_0^1 W_k(x)\, V_n(x)dx$$

(147) $$F_{nk} = 0$$

where

(148) $$\bar{\gamma} = 1 - (1-x)^{\eta}$$
$$\bar{\xi} = (1-x)^{\eta}.$$

With the integrals (147) now known, it is possible to now evaluate the critical flutter load from characteristic equation (117) which may be restated as

(149) $$\text{DET} \mid a_{nk}\lambda^3 + b_{nk}\lambda^2 + c_{nk}\lambda + d_{nk} \mid = 0$$

where

$$a_{nk} = \xi A_{nk}$$

$$b_{nk} = A_{nk} + \xi\bar{q} E_{nk}$$

$$c_{nk} = \gamma B_{nk} - \xi C_{nk} - \xi D_{nk} + \bar{q} E_{nk} + \xi F_{nk}$$

(150) $$d_{nk} = B_{nk} - C_{nk} - D_{nk} + F_{nk}.$$

To allow for the utilization of available numerical procedures, it is convenient at this time to reformulate equation (149) into a standard matrix eigenvalue problem

(151) $$(\underline{R} - \lambda \underline{I})\underline{x} = 0.$$

Clearly, the matrix problem leading to equation (149) is

(152) $$(\underline{A}\,\lambda^3 + \underline{B}\,\lambda^2 + \underline{C}\,\lambda + \underline{D})\underline{x} = 0.$$

If the two definitions

$$\lambda \underline{x} = \underline{I}\,\underline{y}$$

(153) $$\lambda \underline{y} = \underline{I}\,\underline{z}$$

are introduced into equation (152), it is readily shown that the relation

$$\begin{bmatrix} \underline{\Phi} & \underline{I} & \underline{\Phi} \\ \underline{\Phi} & \underline{\Phi} & \underline{I} \\ -\underline{A}^{-1}\underline{D} & -\underline{A}^{-1}\underline{C} & -\underline{A}^{-1}\underline{B} \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{y} \\ \underline{z} \end{bmatrix} = \lambda \begin{bmatrix} \underline{x} \\ \underline{y} \\ \underline{z} \end{bmatrix}$$

(154)

results and is exactly of the form (151).

9. CONSTRAINTS AND THE NUMERICAL PROCEDURE. First of all, two important constraints naturally appear in the consideration of beam optimization. The Euler Beam Theory itself is only valid for a thickness to length ratio less that 1/10th. Since reference will momentarilly be made to comparisons of the minimum design to a fixed rectangular design with $\alpha=\mu$ and since the load intensity as given in equation (62) increases as the clamped end is approached, the clamped end thickness will be taken greater than or at most equal to the free end thickness and hence the constraint

(155) $$\mu \geq \alpha.$$

Hence, since μ is the larger of μ and α, the Euler Beam restriction of a thickness to length ratio less than 1/10 may be expressed as

$$(156) \qquad \mu < \frac{1}{20}$$

as the clamped end thickness is really 2a. Practically, one end of the beam must be of finite thickness or thus μ must be finite, this condition together with equation (156) result in the constraint on μ that

$$(157) \qquad 0 < \mu < \frac{1}{20} .$$

Now, with equation (157) in mind, the problem of avoiding negative areas is accomplished by setting the free end thickness to being non-negative which together with equation (155) results in the constraint on α being

$$(158) \qquad 0 \leq \alpha \leq \mu .$$

Additionally, the calculations for an optimum design will be related to a fixed standard of a rectangular beam with

$$(159) \qquad \alpha = \mu$$

and its corresponding flutter load

$$(160) \qquad Q_{\substack{crit. \\ rect.}} = \frac{3}{4} \frac{q\ell}{Eb} .$$

Thus, the method of fixing the flutter load will also include determining an equivalent rectangular beam with the same prespecified flutter load and using it and its dimensions as a reference. In terms of this rectangular reference beam and realizing that equation (158) must hold, an added restriction upon μ is that

$$(161) \qquad \mu_{rect.} \leq \mu$$

where $\mu_{rect.}$ is now the value of μ for the rectangular reference beam. Thus, from equations (157), (158), and (161) the constraints to be used in the optimization routine are

$$(162) \qquad \mu_{rect.} \leq \mu < \frac{1}{20}$$

$$(163) \qquad 0 \leq \alpha \leq \mu .$$

10. DISCUSSION. As a preliminary to the mass optimization work, some results will now be shown for the stability of Hauger's problem. The various effects of internal and external damping on flutter load will be carefully considered.

Figure 3 shows the variation of external damping, $\bar{q}$, versus flutter load, Q, for various fixed values of internal damping, ξ, with internal damping γ=0. For each fixed value of internal damping, ξ, the general tendency is for flutter load to increase as external damping, $\bar{q}$, increases until a flutter load of approximately $Q \times 10^4 = 11.0$ is reached; at this time there is a decrease in flutter load as external damping, $\bar{q}$, further increases with an asymptotic value of $Q \times 10^4 = 8.3$ eventually being reached. Notice that as the fixed values of internal damping, ξ, increase that the values of external damping, $\bar{q}$, at which flutter load reaches a maximum also increase. Also note that from this last result it would appear that for a given value of external damping, $\bar{q}$, a maximum flutter load can be obtained by making a proper choice of internal damping, ξ.

Figure 4 now shows the variation of internal damping, ξ, versus flutter load, Q, with external damping, $\bar{q}$, now taking on fixed values -- the other internal damping, γ, is again zero. As internal damping, ξ, increases, flutter load begins at a value of $Q \times 10^4 = 8.3$ and increases to a maximum of approximately $Q \times 10^4 = 11.0$ and then continually decreases for each of the fixed values of external damping, $\bar{q}$. As fixed values of external damping, $\bar{q}$, increase, the values of internal damping, ξ, at which flutter load reaches a maximum also increase. For a given value of internal damping, ξ, a maximum flutter load can be obtained by making a proper choice of external damping, $\bar{q}$.

Figure 5 shows the variation of flutter load, Q, with internal damping, γ, for fixed values of internal damping, ξ, for external damping set at $\bar{q} = 1.0$. For some of the higher values of fixed values of internal damping, ξ, there is a general tendency for flutter load to increase as internal damping, γ, increases to a maximum in the range $10.8 < Q \times 10^4 < 11.4$ and then to gradually decrease. For some of the lower values of internal damping, ξ, flutter load starts out at its maximum at internal damping γ=0.0 and then decreases as internal damping, γ, increases. A maximum flutter load for a pre-specified internal damping, γ, can be obtained through a proper choice of internal damping, ξ. As fixed values of internal damping, ξ, increase, the values of internal damping, γ, at which flutter load reaches a maximum also increase, but unlike the trend in figures 3 and 4, the value of this maximum tends to vary slightly.

Figure 6 shows the variation of flutter load, Q, with internal damping, ξ, for fixed values of internal damping, γ, with external damping set at $\bar{q} = 1.0$. For the various fixed values of internal damping, γ, it is clear that flutter load increases as internal damping, ξ, increases up to a maximum of $Q \times 10^4 = 11.4$ or less and then flutter load decreases with further increases in internal damping, ξ. As fixed values of internal damping, γ, increase, the values of internal damping, ξ, at which flutter load reaches a maximum also increases, but the value of the maximum varies somewhat.

Figure 7 shows the variation of base thickness μ with tip thickness α for internal dampings $\gamma = 1000.$ and $\xi = 1.0$ and for external damping $\bar{q} = 1.0$ where the curves themselves are for various fixed values of η, the order of the generalized parabola as given in equation (132). In Figure 7, the constraint $\mu < .05$ as expressed in equation (156) is utilized and will be held to in all future plots. Curves for $\eta = 1,2,3,4$ are shown and it is clear that base thickness μ increases as order η increases or as tip thickness α decreases. The results of Figure 7 may be directly applied to calculate the mass ratio

$$(164) \qquad M/M_0 = \alpha + (\mu-\alpha)/(\eta+1)$$

where

$$(165) \qquad M_0 = 2 \rho \ell^2 b$$

as equations (164) and (165) directly result from equations (135) and (138). Figure 8 shows the variation of mass ratio M/M_0 for generalized parabola's of order $\eta = 1,2,3,4$. It is clear that the minimum mass ratio will occur for zero tip thickness, $\alpha = 0$, which would yield a beam with a knife edge at the tip. Thus, in any practical applications it would be wise to specify a minimum tip thickness as a lower constraint on α. It is further clear from figure 8 that different orders of generalized parabolas yield the lower value of mass ratio M/M_0 for different values of α. For example, $\eta = 2$ gives the lowest mass ratio in the range $0 \leq \alpha \leq .00825$, $\eta = 3$ gives the lowest mass ratio in the range $.00825 \leq \alpha \leq .00915$, and $\eta = 4$ gives the lowest mass ratio in the range $.00915 \leq \alpha \leq .0167$. It is also clear that if a beam with a knife edge were allowed, a beam with a mass reduction in excess of 24.5% would be possible.

Now, as an example of a true mass optimization with a lower constraint on tip thickness the beam represented in figures 7 and 8 will be optimized with a Rosenbrock algorithm [9]. For a lower bound of $\alpha = .000833$ the Rosenbrock algorithm yields an optimum design at this value of α of a mass ratio of $M/M_0 = .766$ corresponding to a base thickness $\mu = .0378$ for a generalized parabola of order $\eta = 2.1$. Thus, for this constrained problem a mass reduction of 23.4% is possible.

Figure 9 shows the variation of base thickness μ with tip thickness α for internal dampings $\gamma = 100.$ and $\xi = 1.0$ and for external damping $\bar{q} = 1.0$ for generalized parabolas of order $\eta = 1,2,3,4$. Figure 10 shows the corresponding variation of mass ratio M/M_0 with tip thickness α. Clearly, $\eta = 2$ gives the lowest mass ratio in the range $0 \leq \alpha < .00565$, $\eta = 3$ gives the lowest mass ratio in the range $.00565 \leq \alpha \leq .00745$, and $\eta = 4$ gives the lowest mass ratio in the range $.00745 \leq \alpha \leq .0167$. Notice that for low values of α the curves for $\eta = 3$ and $\eta = 4$ give much higher mass ratios and then from figure 9 eventually overshoot the constraint that $\mu < .05$. A beam with a knife edge would result in a mass reduction of at least 27.5%; in this case a beam with a finite tip thickness would result in a better mass reduction of at least 27.6% and indicates at least one case where the knife edge condition is not the minimum.

For a lower bound of $\alpha = .000833$ the Rosenbrock algorithm yields an optimum design at this value of α of a mass ratio of $M/M_0 = .709$ corresponding to a base thickness $\mu = .0308$ for a generalized parabola of order $\eta = 1.73$ and hence a mass reduction of 29.1% for this constrained problem.

Figure 11 shows the variation of base thickness μ with tip thickness α for internal dampings $\gamma = 1000$ and $\xi = 10.0$ and for external damping $\bar{q} = 1.0$ for generalized parabolas of order $\eta = 1,2,3,4$. Figure 12 shows the corresponding variation of mass ratio M/M_0 with tip thickness α for the parameters specified in figure 11. Note that $\eta = 2$ gives the lowest mass ratio in the range $0 \leq \alpha \leq .0083$, $\eta = 3$ gives the lowest mass ratio in the range $.0083 \leq \alpha \leq .0094$, and $\eta = 4$ gives the lowest mass ratio in the range $.0094 \leq \alpha \leq .0167$. In this instance a beam with a knife edge would result in a mass reduction of at least 24.8%. For a lower bound of $\alpha = .000833$ the Rosenbrock algorithm yields an optimum design at this value of α of a mass ratio of $M/M_0 = .765$ for a generalized parabola of order $\eta = 2.1$ and hence a mass reduction of 23.5% for this constrained problem.

Figure 13 shows the variation of base thickness μ with tip thickness α for internal dampings $\gamma = 1000$ and $\xi = 10.0$ and an external damping of $\bar{q} = 10.0$ for generalized parabolas of order $\eta = 1,2,3,4$. Figure 14 shows the corresponding variation of mass ratio M/M_0 with tip thickness α. Clearly $\eta = 2$ gives the lowest mass ratio in the range $0 \leq \alpha \leq .0057$, $\eta = 3$ gives the lowest mass ratio in the range $.0057 \leq \alpha \leq .0074$, and $\eta = 4$ gives the lowest mass ratio in the range $.0074 \leq \alpha \leq .0167$. A beam with a knife edge would result in a mass reduction of at least 27.4% with slightly lower values possible for a blunt end. For a lower bound of $\alpha = .000833$ the Rosenbrock algorithm yields an optimum design at this value of α of a mass ratio of $M/M_0 = .708$ corresponding to a base thickness of $\mu = .0309$ for a generalized parabola of order $\eta = 1.726$ and hence a mass reduction of 29.2% for this constrained problem.

Figure 15 shows the variation of base thickness μ with tip thickness α for internal dampings $\gamma = 500$ and $\xi = 10.0$ and an external damping of $\bar{q} = 10.0$ for generalized parabolas of order $\eta = 1,2,3,4$. Figure 16 shows the corresponding variation of mass ratio M/M_0 with tip thickness α. In this case $\eta = 2$ gives the lowest mass ratio in the range $0 \leq \alpha \leq .00475$, $\eta = 3$ gives the lowest mass ratio in the range $.00475 \leq \alpha \leq .0065$, and $\eta = 4$ gives the lowest mass ratio in the range $.00665 \leq \alpha \leq .0167$. A beam with a knife edge would result in a mass reduction of at least 31.5% and one with a tip thickness of $\alpha = .0004$ would result in a mass reduction of at least 32%, but this is still a very small tip thickness. For a lower bound of $\alpha = .000833$ the Rosenbrock algorithm yields an optimum design at this value of α of a mass ratio of $M/M_0 = .682$ corresponding to a base thickness $\mu = .0325$ for a generalized parabola of order $\eta = 1.97$ and hence a mass reduction of 31.8% for this constrained problem.

TABLE 1. SUMMARY OF ROSENBROCK OPTIMIZATION RESULTS

CASE	γ	ξ	$\bar{q}$	$\alpha x 10^4$	μ	η	M/M_o	% DECREASE
1	1000	1.0	1.0	8.33	.0378	2.1	.766	23.4
2	100	1.0	1.0	8.33	.0308	1.73	.709	29.1
3	1000	10.0	1.0	8.33	.0378	2.1	.765	23.5
4	1000	10.0	10.0	8.33	.0309	1.726	.708	29.2
5	500	10.0	10.0	8.33	.0325	1.97	.682	31.8

Table 1 summarizes the results of the Rosenbrock optimizations that were previously discussed in figures 7 thru 16. It is clear that from the five sets of parameters for which an optimization was made that reductions in mass of from 23.4% to 31.8% are possible with a tip thickness constraint of .000833 specified.

11. CONCLUSIONS. An adjoint variational principle has been applied to the flutter stability problem of Euler beams with both internal and external damping. An application of a generalized Ritz approximation and the application of a variational principle resulted in a characteristic equation for flutter load. Hauger's problem was then optimized within the bounds of a special class of parabolic shape functions.

Preliminary studies of the variation of internal dampings and external damping for the stability problem indicated suitable ranges of these parameters for use in the optimization problem. The various plots of mass ratio versus tip thickness yielded a vital insight into what should be expected from an optimization routine. It was quite clear that a minimum mass ratio was obtainable in each case for a knife edge at the beams free end, but this condition is impractical. Thus, a minimum tip thickness constraint must be applied to any optimization problem to obtain realistic results. In going to a Rosenbrock optimization routine with a minimum tip constraint of $\alpha = .000833$ it is quite clear that mass ratio reductions of anywhere from 23.4% to 31.8% are possible.

REFERENCES:

1. WILFRED E. BAKER, WILLIAM E. WOOLAM, and DANA YOUNG (1967). Air and Internal Damping of Thin Cantilever Beams. Int. J. Mech. Sci., 9, 743-766.

2. E. J. BRUNELLE (1970). J. Composite Materials, 4, 404-416. The Statics and Dynamics of a Transversely Isotropic Timoshenko Beam.

3. E. J. BRUNELLE (1970). AIAA Journal, 8, 2271-2273. Elastic Instability of Transversely Isotropic Timoshenko Beams.

4. MAX BECK (1952). Zeit. Angew. Math. Phys., 3, 225-228. Die knicklast des einseitig eingespannten, tangental gedrückten stabes.

5. HORST LEIPHOLZ (1962). Zeit. Angew. Math. Phys., 13, 581-589. Die knicklast des einseitig eingespannten stabes mit gleichmässig verteilter, tangentialer längsbelastung.

6. W. HAUGER (1966). Ingen.-Arch., 35, 221-229. Die knicklasten elastischer stäbe unter gleichmässig verteilten und linear veränderlichen, tangentialen druckkräften.

7. G. L. ANDERSON (1973). Journal of Sound and Vibration, 27, 279-296. Application of a Variational Method to Dissipative, Non-Conservative Problems of Elastic Stability.

8. C. R. THOMAS (1973). Watervliet Arsenal Technical Report R-WV-T-6-26-73. A Relative Optimization of Cantilever Euler Beams with Example For Hauger's Problem.

9. H. H. ROSENBROCK (1960). Computer Journal 3, 175-184. An Automatic Method for Finding the Greatest or Least Value of a Function.

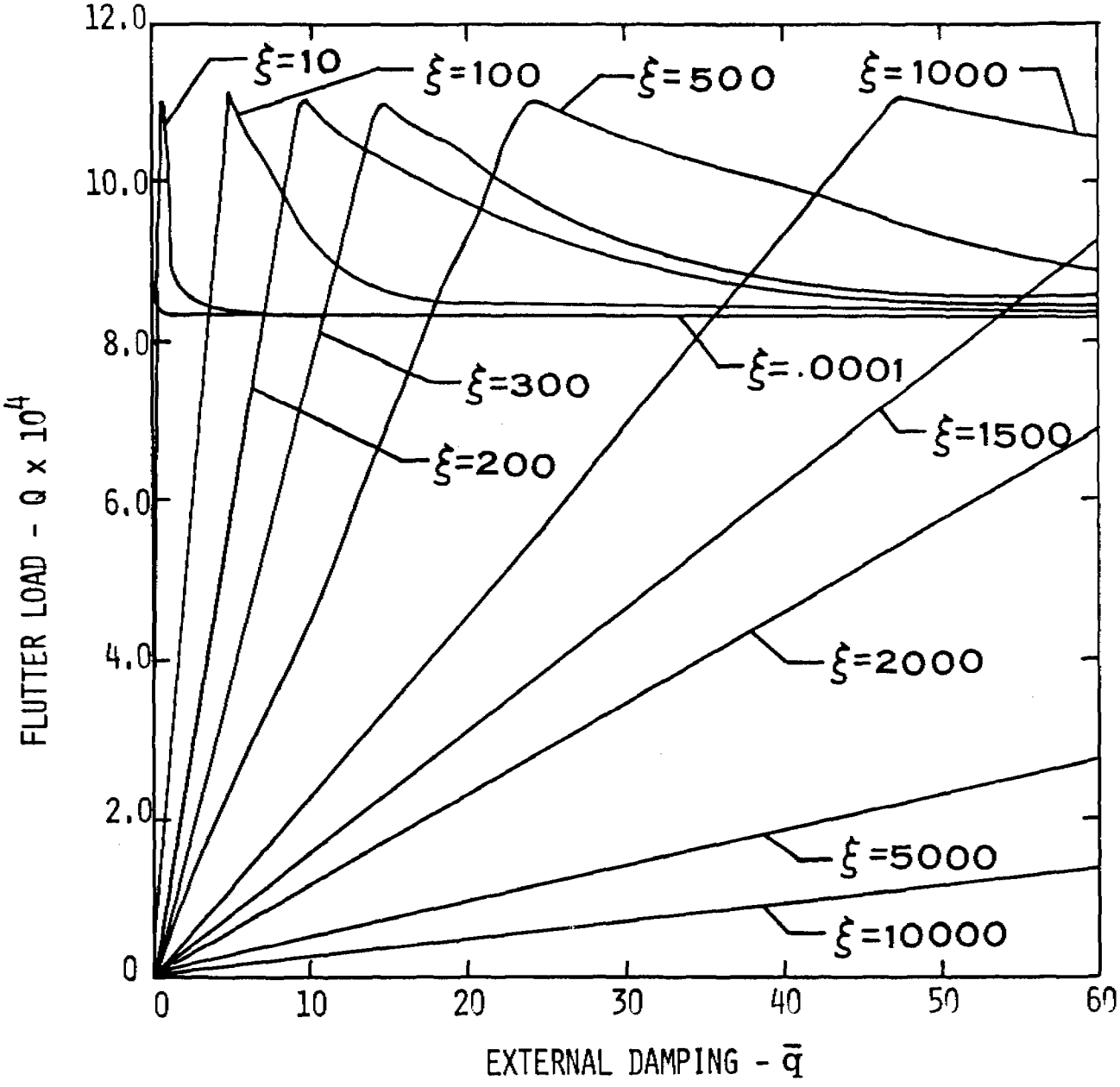


Figure 3. Variation of Flutter Load Q with External Damping $\bar{q}$ for Fixed Values of Interval Damping ξ with $\gamma=0.0$ for Uniform Cross-Section

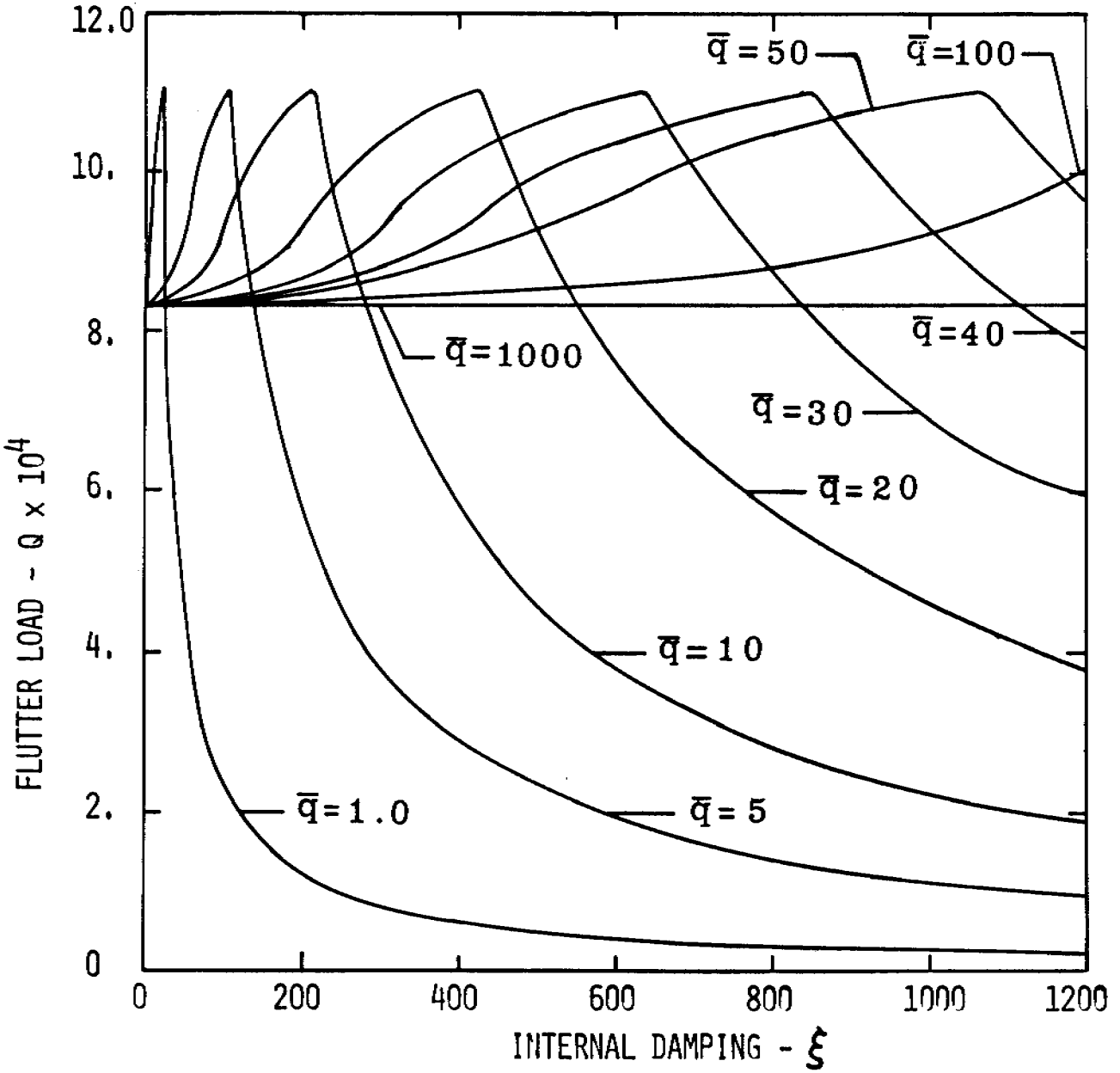


Figure 4. Variation of Flutter Load Q with Internal Damping ξ for Fixed Values of External Damping $\bar{q}$ with γ=0.0 for Uniform Cross-Section.

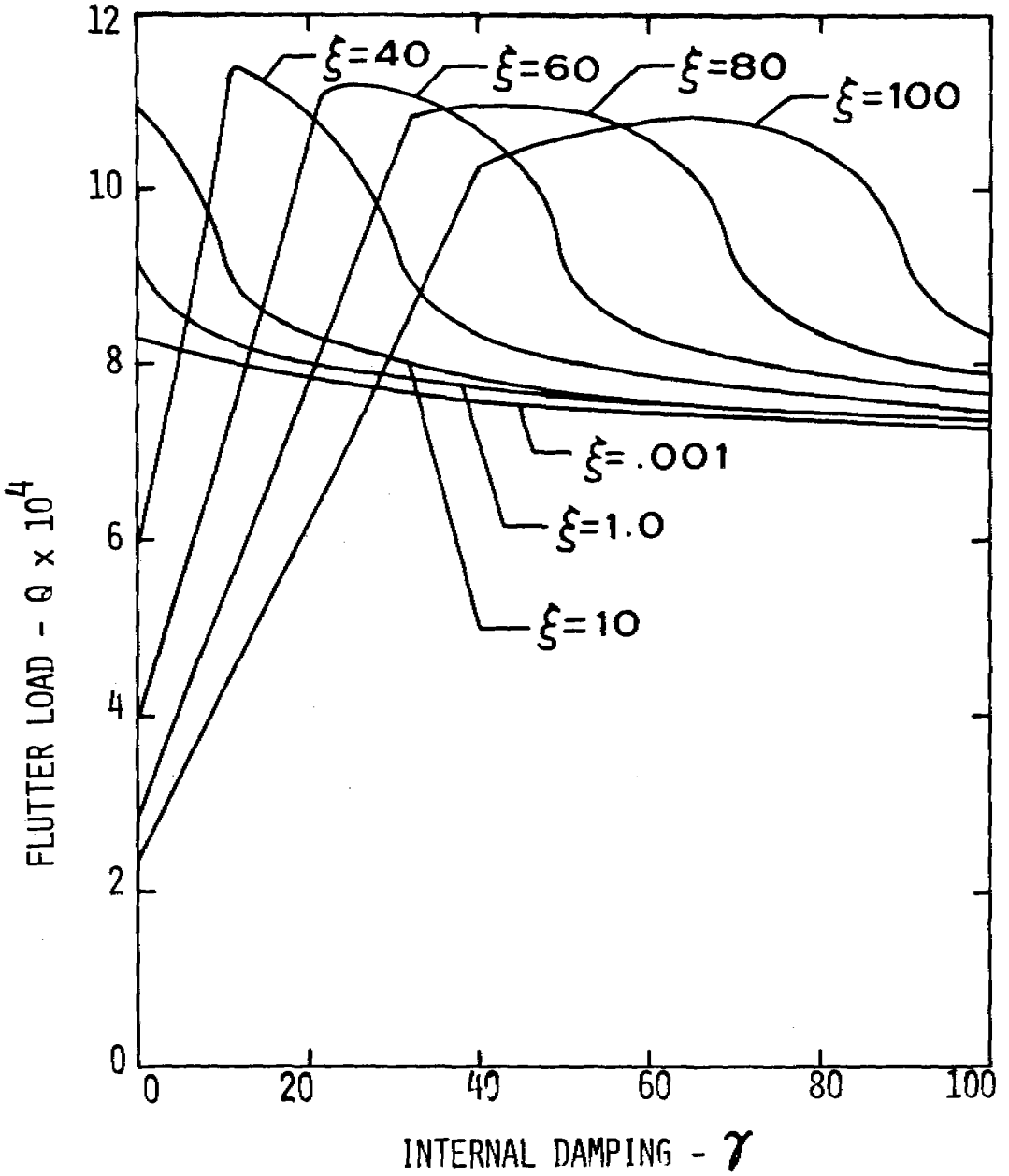


Figure 5. Variation of Flutter Load Q with Internal Damping γ for Fixed Values of Internal Damping ξ with $\bar{q}$=1.0 for Uniform Cross-Section.

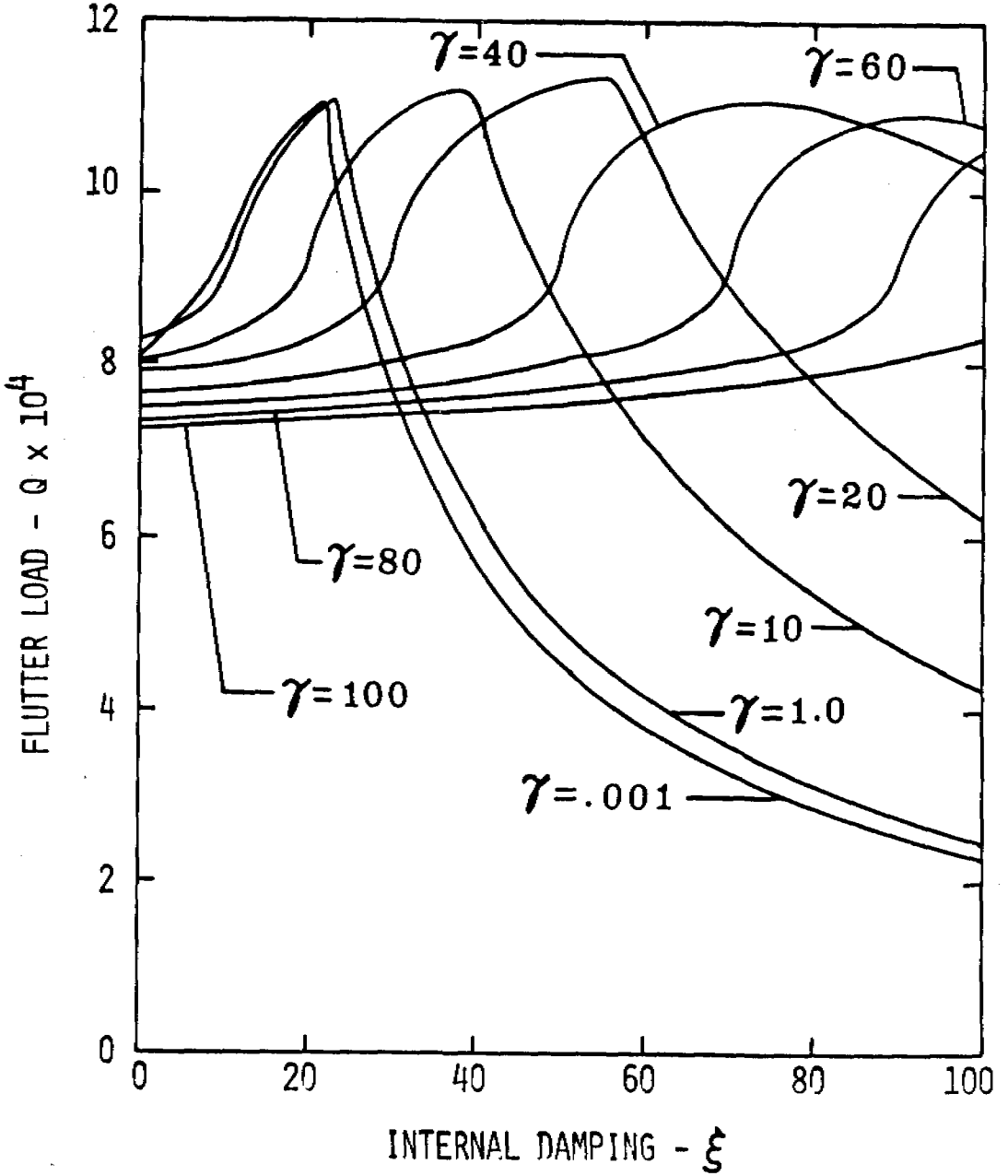


Figure 6. Variation of Flutter Load Q with Internal Damping ξ for Fixed Values of Internal Damping γ with $\bar{q}$=1.0 For Uniform Cross-Section.

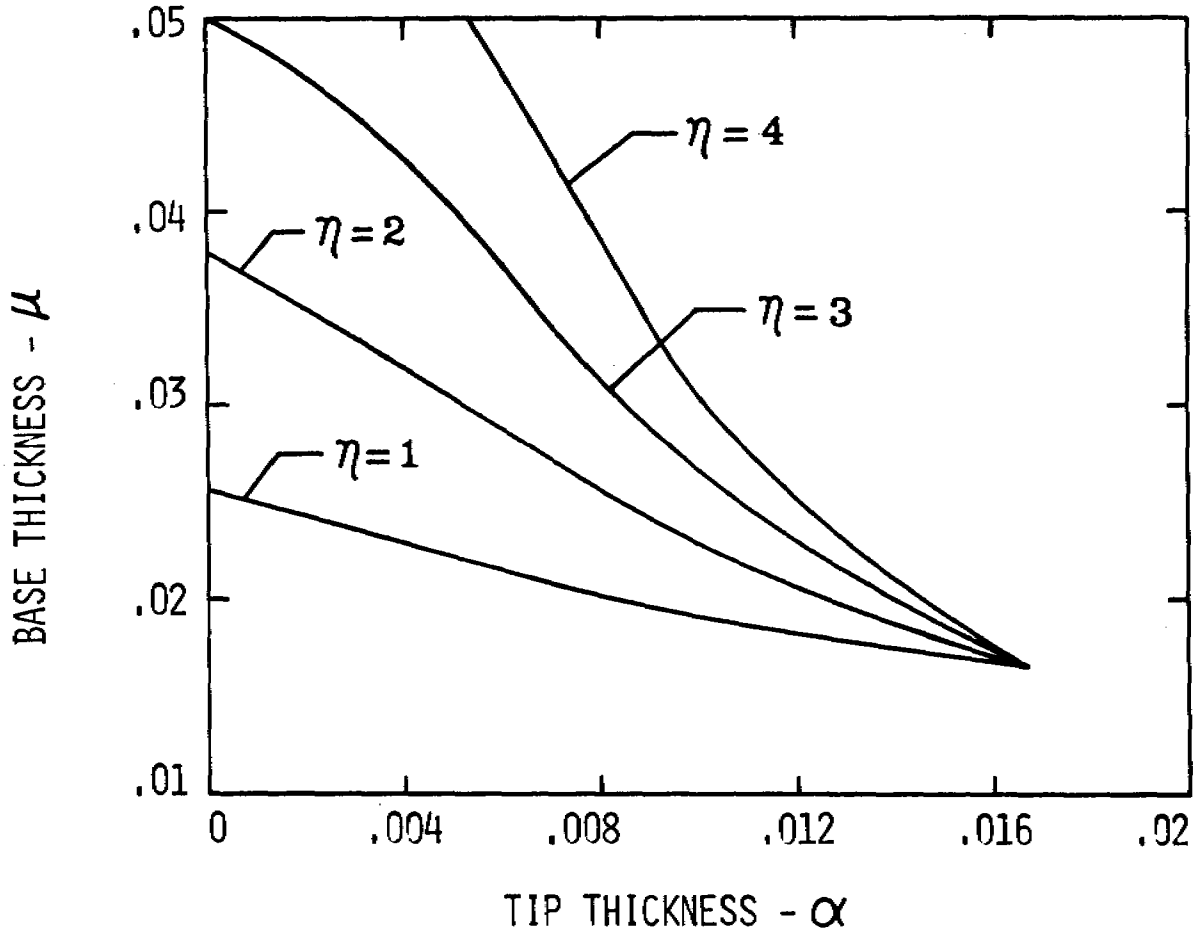


Figure 7. Variation of Base Thickness with Tip Thickness for γ=1000, ξ=1.0, and $\bar{q}$=1.0.

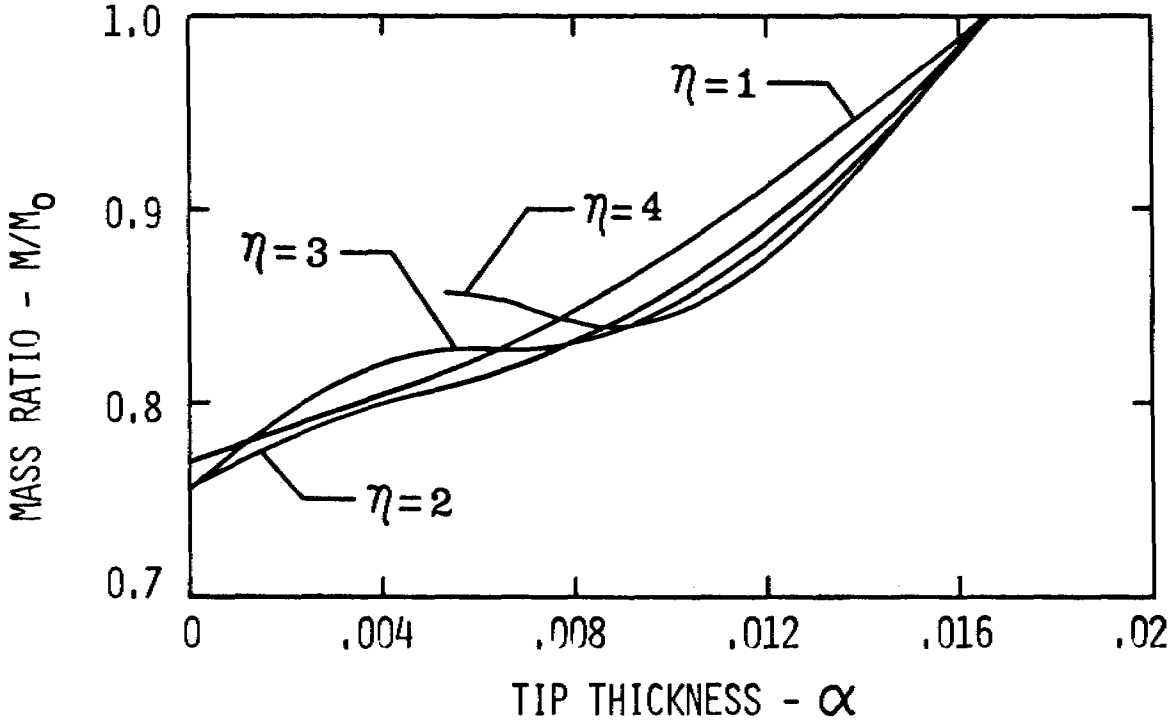


Figure 8. Variation of Mass Ratio with Tip Thickness for $\gamma=1000$, $\xi=1.0$, and $\bar{q}=1.0$.

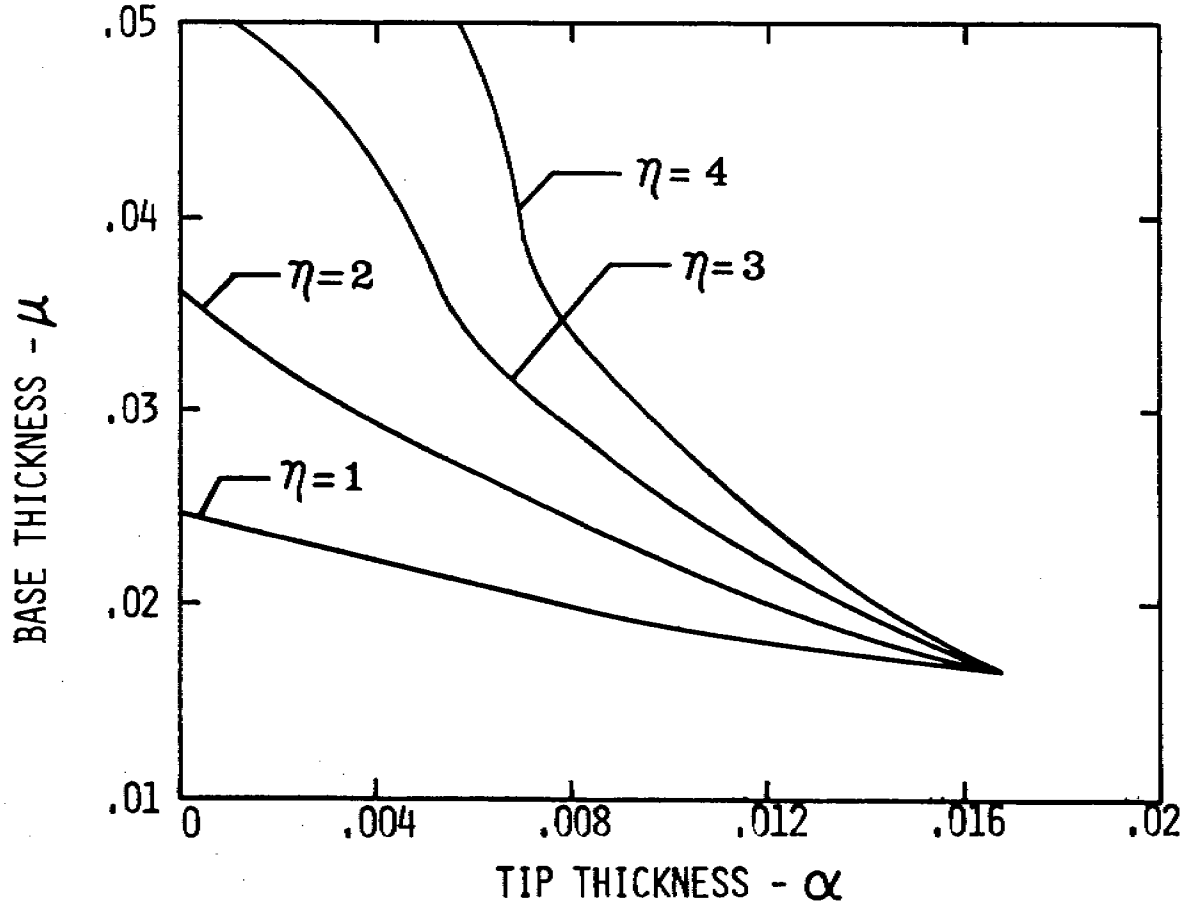


Figure 9. Variation of Base Thickness with Tip Thickness for γ=100., ξ=1.0, and $\bar{q}$=1.0.

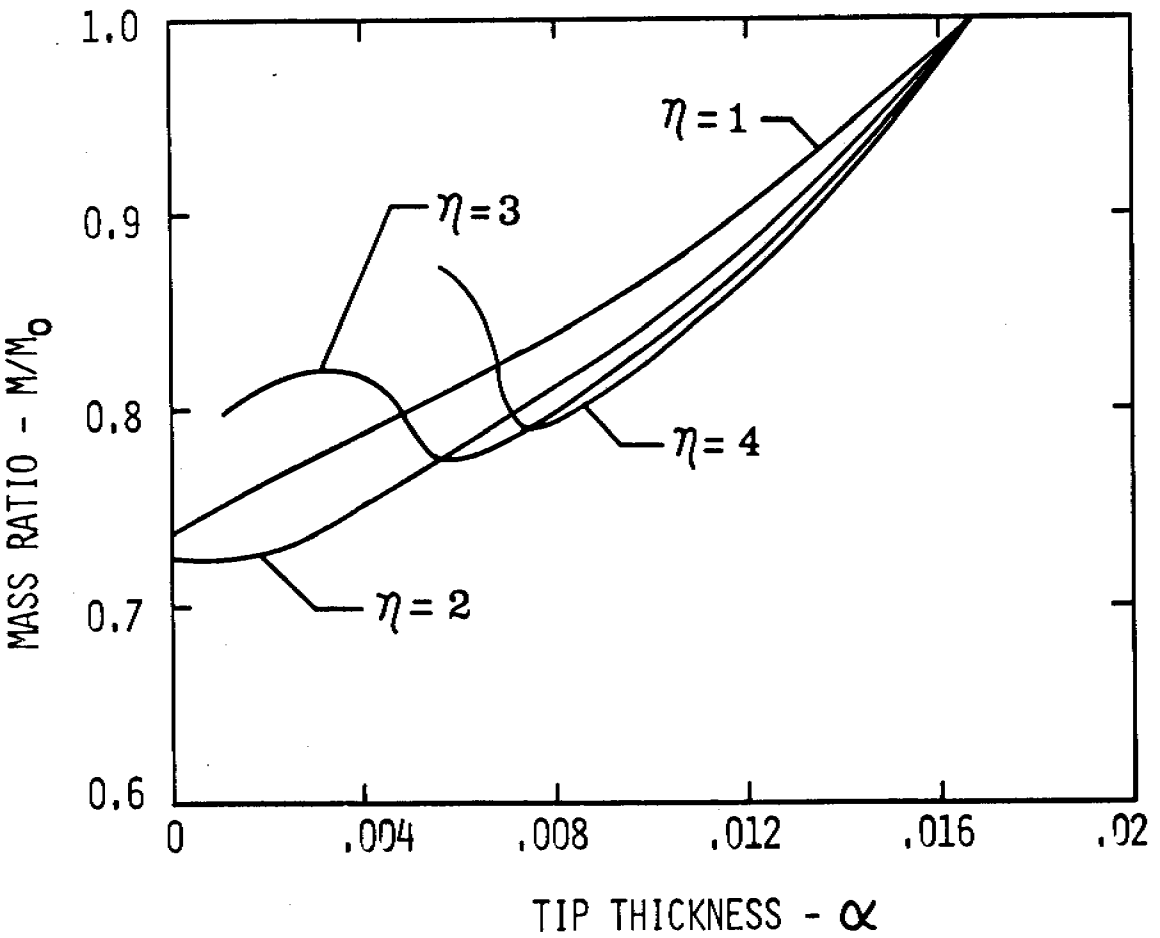


Figure 10. Variation for Mass Ratio with Tip Thickness for γ=100., ξ=1.0, and $\bar{q}$=1.0.

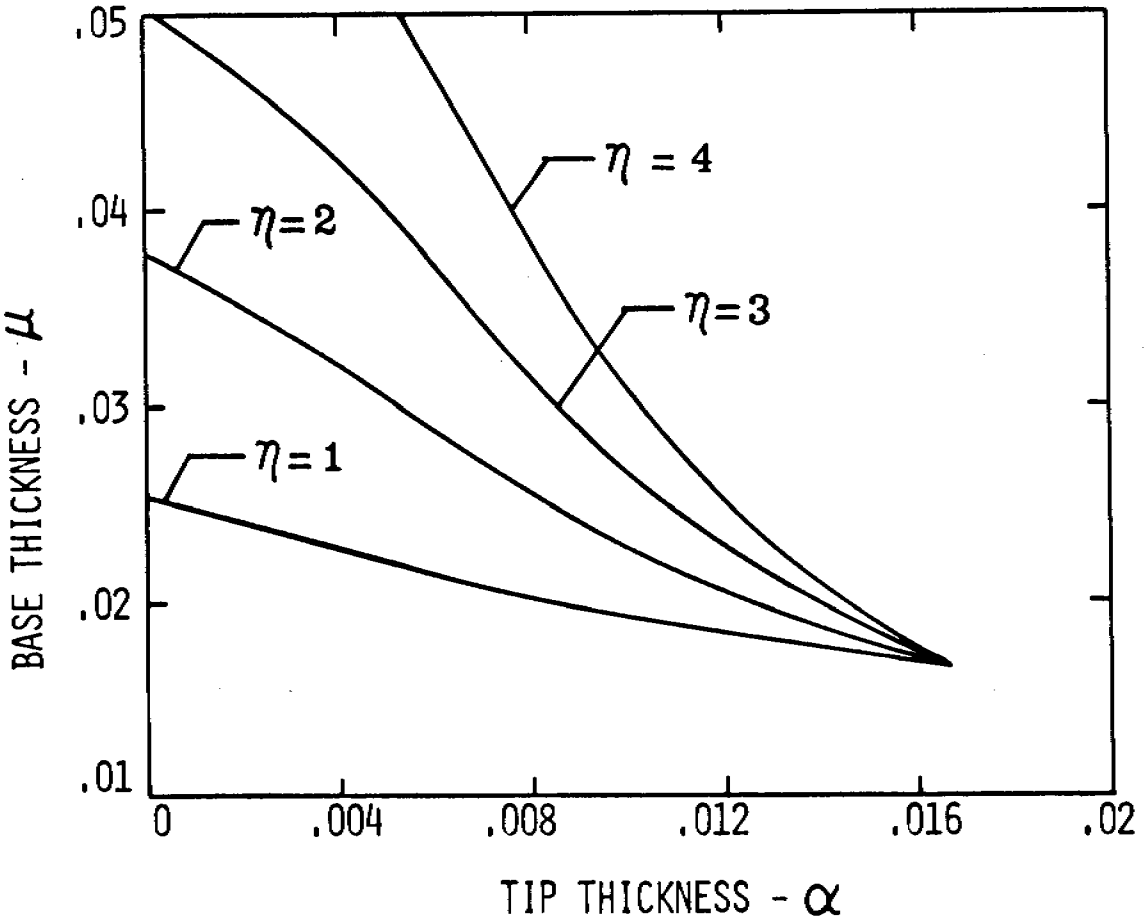


Figure 11. Variation of Base Thickness with Tip Thickness for γ=1000, ξ=10., and $\bar{q}$=1.0.

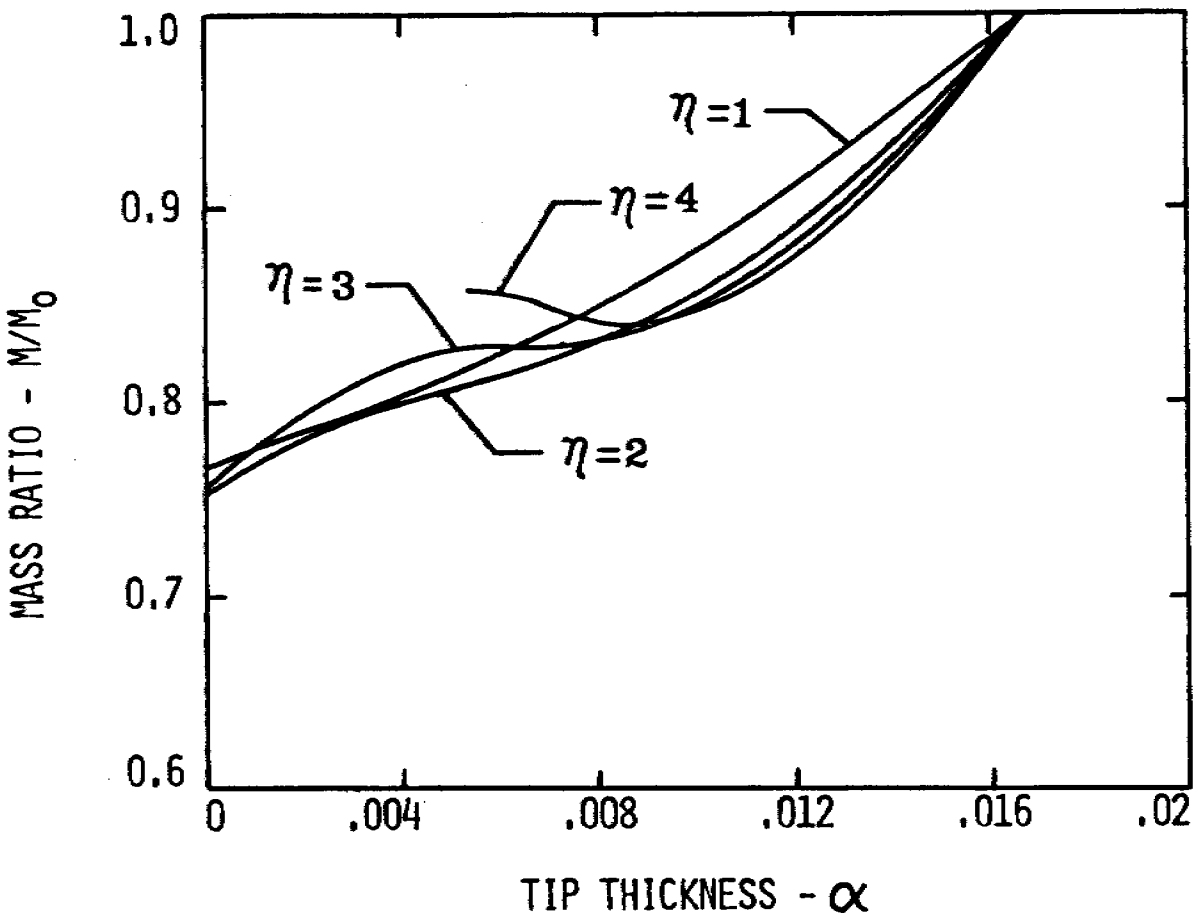


Figure 12. Variation of Mass Ratio with Tip Thickness for γ=1000, ξ=10., and $\bar{q}$=1.0.

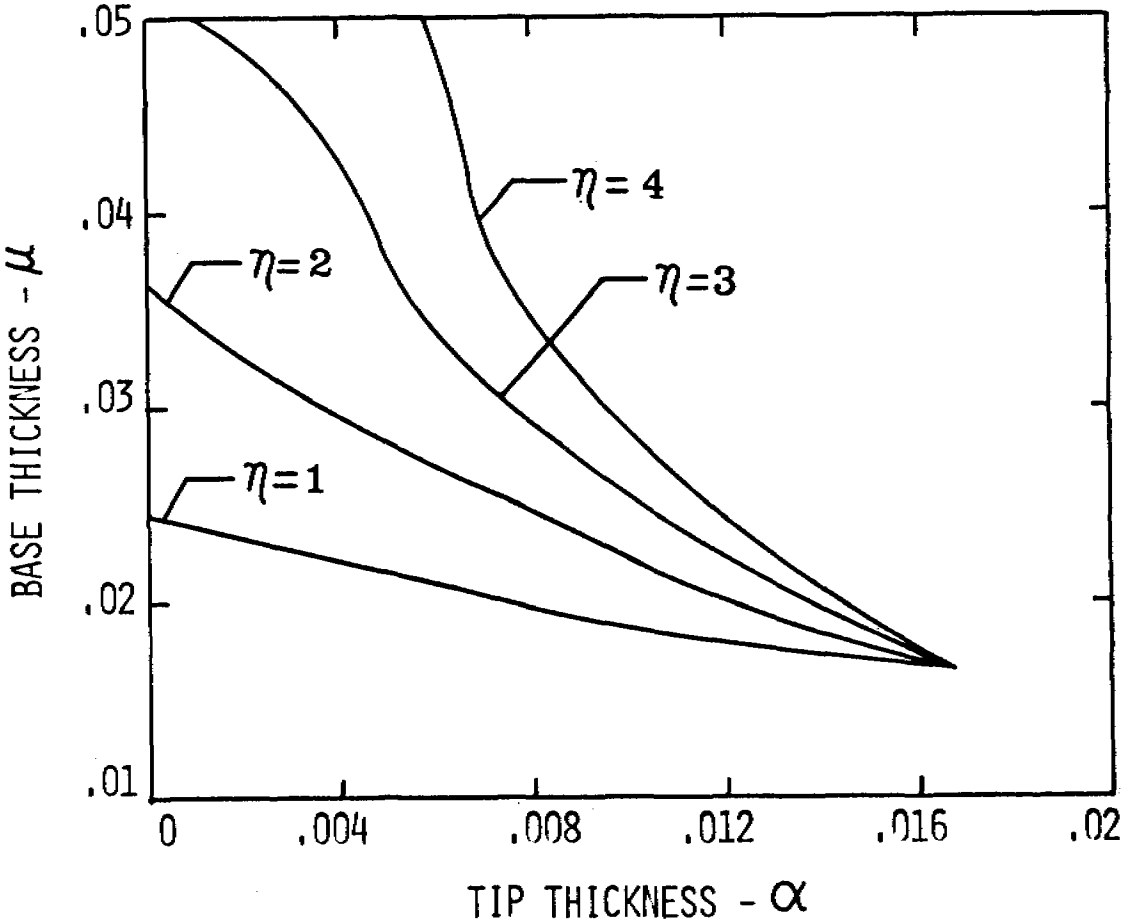


Figure 13. Variation of Base Thickness with Tip Thickness for γ=1000, ξ=10., and $\bar{q}$=10.

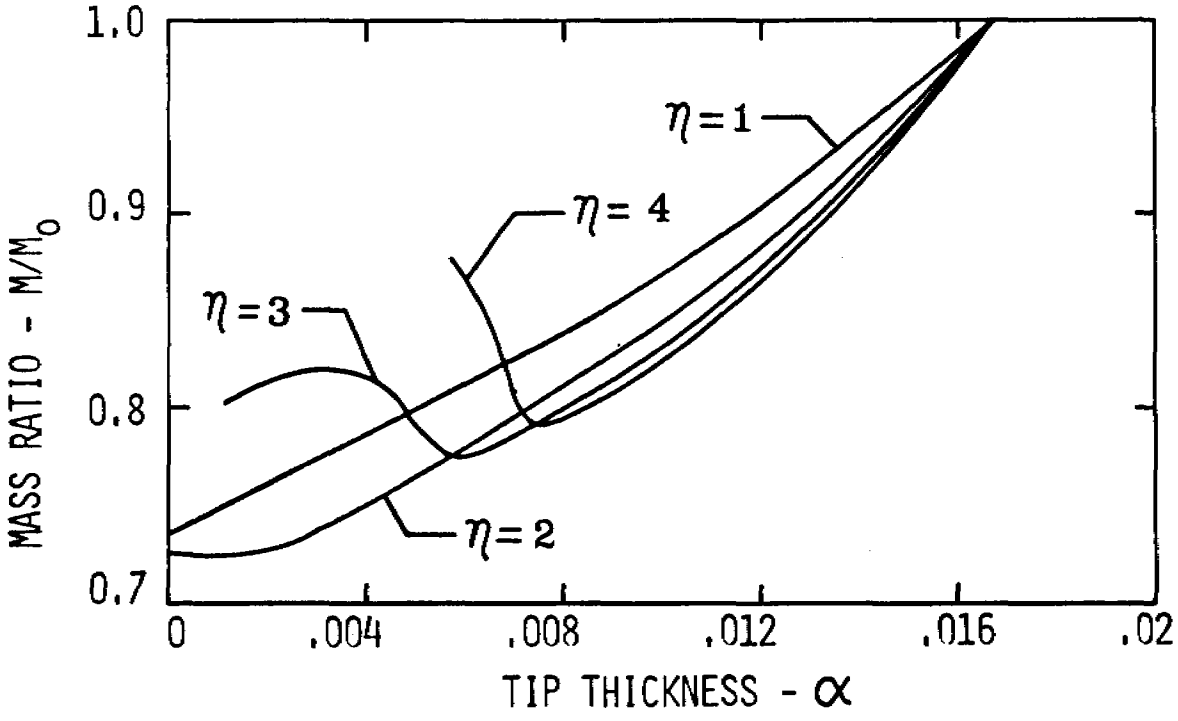


Figure 14. Variation of Mass Ratio with Tip Thickness for γ=1000, ξ=10., and $\bar{q}$=10.

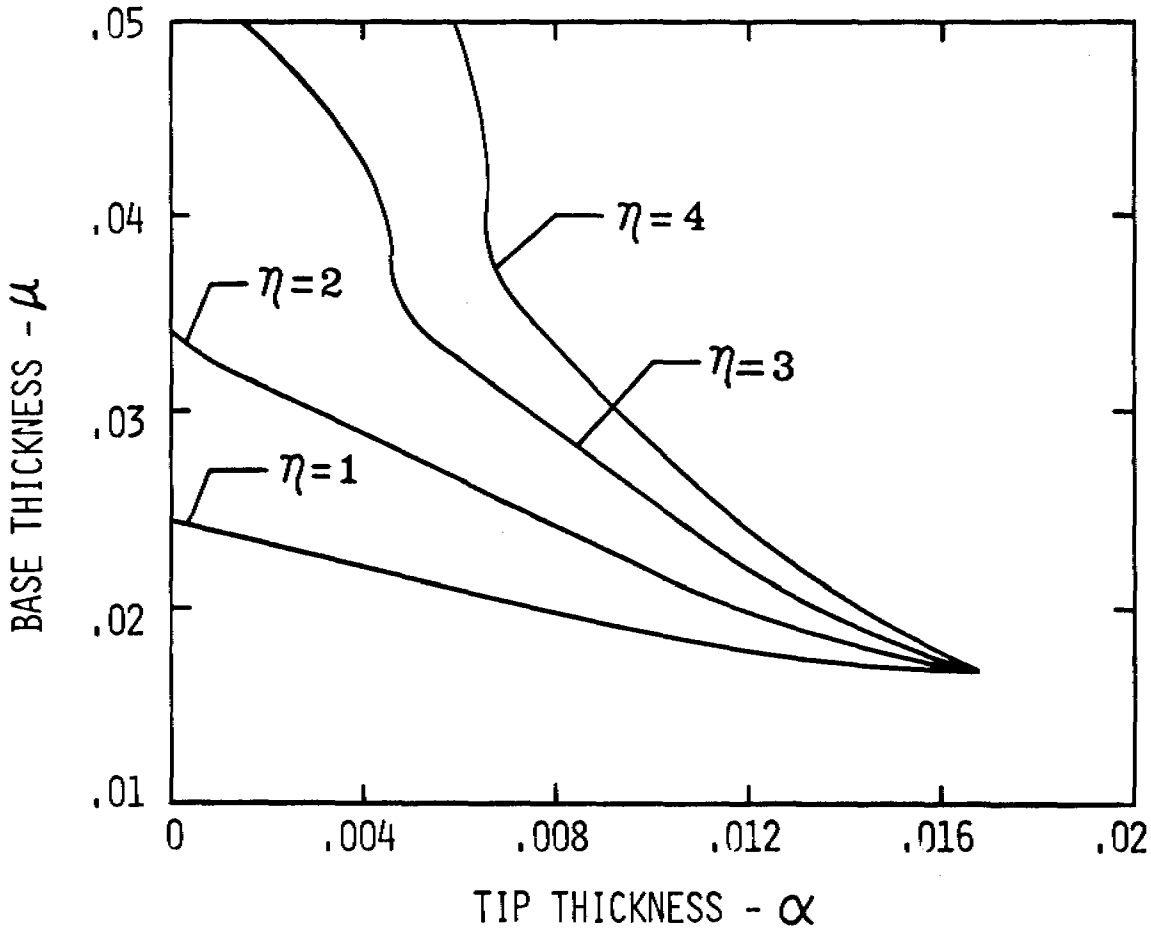


Figure 15. Variation of Base Thickness with Tip Thickness for γ=500, ξ=10., and $\bar{q}$=10.

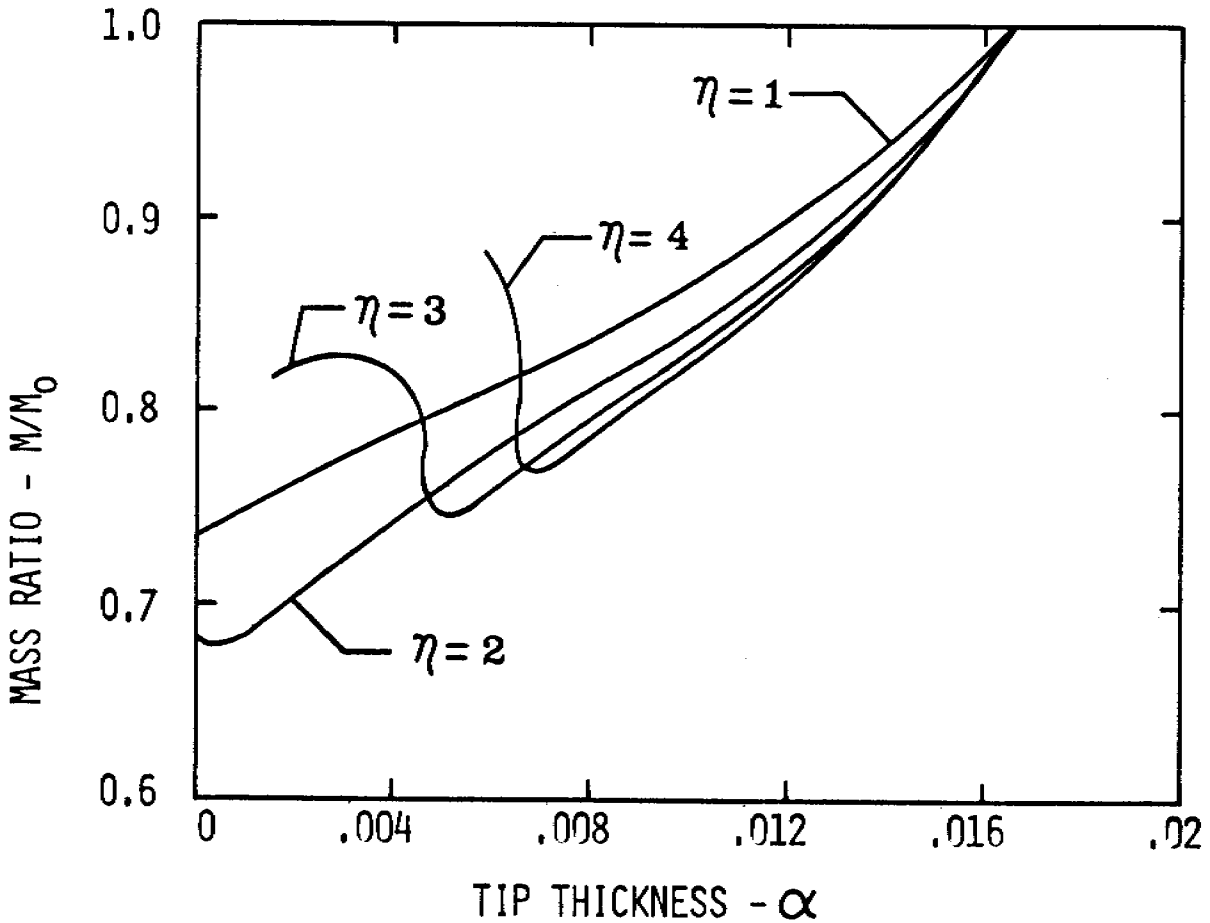


Figure 16. Variation of Mass Ratio with Tip Thickness for γ=500, ξ=10., and $\bar{q}$=10.

Buckling of Orthotropic Rectangular Cylinders*

Earl C. Steeves

U.S. Army Natick Laboratories

Natick, Massachusetts

One of the loading situations that must be considered in the design of packages is that resulting from stacking. Since this is a compressive loading, the buckling mode of failure must be considered. To obtain some guidance for this design problem, solutions to the buckling problem for rectangular cylinders were obtained. These solutions were obtained by direct solution of the differential equations and by approximate solution using the Rayleigh-Ritz procedure. Both solutions are based on classical orthotropic plate theory and treat the corners by conditions of continuity. The direct solution which is restricted to the uniform load results in a transcendental eigenvalue problem for the buckling load and is solved by a search and interpolation scheme. The approximate solution utilizes finite difference approximations in conjunction with the Rayleigh-Ritz analysis and is applicable to both the uniform and nonuniform load situations. The constraint equations describing the corner conditions are incorporated in the Rayleigh-Ritz procedure by a technique not commonly employed. The resulting algebraic eigenvalue problem is solved using the Jacobi reduction procedure. Results from the direct and approximate solutions are compared for the uniform load case and results for the nonuniform load case obtained using the approximate solution are compared with experimental results for rectangular fiberboard cylinders. These comparisons with experiment indicate a need to examine the nonlinear buckling problem and to include plasticity effects in the case of fiberboard.

---

*This paper will be published as U.S. Army Natick Development Center Technical Report 75-99 entitled, "Theoretical and Experimental Investigation of the Buckling of Rectangular Cylinders".

ENERGY RELEASE RATE IN TERMS OF COMPLEX ANALYTIC FUNCTIONS

M. A. Hussain and S. L. Pu
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. The energy release rate $\mathcal{G}$ for two dimensional cracks subjected to either plane loading or anti-plane shear is most commonly computed by Irwin's crack closure method or the path independent integrals used by Rice. In terms of analytic functions of a complex variable, these two approaches give two distinct expressions for the energy release rate. For instance, for plane problem the energy release rate is given in terms of one analytic function using Irwin's approach while the same is expressed in terms of two analytic functions by the method of path independent integrals. For special cases, such as Mode I crack, these two expressions lead to identical results. However, the identity for the general case has not been established. In this report we have shown that the expression of $\mathcal{G}$ for plane problem in terms of two analytic functions can be further reduced to the same expression obtained by Irwin's method in terms of only one analytic function and that the two distinct expressions of $\mathcal{G}$ for anti-plane problems are identical.

1. INTRODUCTION. The energy release rate $\mathcal{G}$, for two dimensional crack problems subjected to either plane loading conditions or longitudinal shear can be computed by various methods. The most common ones are the 'crack closure' approach of Irwin [1][1] and path independent integrals used by Rice [2]. The equivalence of these two methods is rigorously derived in [2], [3], and [4].

Such crack problems are usually formulated in terms of one or two analytic functions of a complex variable. Two distinct expressions for $\mathcal{G}$ are obtained in terms of analytic functions by these two methods. For self similar crack extensions for which these analytic functions can be explicitly derived, these two different approaches, without any ambiguity, lead to identical results. However for non self similar situations, for example in branching of a crack, one must be very careful in regard to the use of path independent integrals in calculating $\mathcal{G}$. In such a case it is necessary to evaluate the path independent integrals around a contour just surrounding the tip of the branched crack and not around the entire end of crack involving the main crack

[1]Numbers in brackets designate References at end of the paper.

and the branched crack, (see e.g. [5]). Using this approach it is shown in this paper that the path independent integrals and the "crack closure" approach lead to identical results in terms of analytic functions. This equivalence is shown by the use of mapping techniques.

2. COMPUTATION OF $\mathcal{G}$ USING IRWIN'S APPROACH. Considering two analytic functions $\phi_1(z)$, $\psi_1(z)$ of a complex variable $z=x+iy$, the stresses and displacement for the plane problem can be written as [6]:

$$\sigma_x + \sigma_y = 4\,\mathrm{Re}\{\phi_1'(z)\} \tag{1}$$

$$\sigma_y - \sigma_x + 2i\,\tau_{xy} = 2(\bar{z}\phi_1''(z) + \psi_1'(z)) \tag{2}$$

$$\begin{aligned} 2\mu(u+iv) &= \kappa\,\phi_1(z) - z\,\overline{\phi_1'(z)} - \overline{\psi_1(z)} \\ i\int (X_n + iY_n)ds &= \phi_1(z) + z\,\overline{\phi_1'(z)} + \overline{\psi_1(z)} \end{aligned} \tag{3}$$

For the anti-plane problem, using $F(z)$ as an analytic function, we have the stresses and displacement :

$$\tau_{xz} - i\,\tau_{yz} = \mu\,F'(z) \tag{4}$$

$$w = \mathrm{Re}\{F(z)\} \tag{5}$$

It is convenient to use a mapping function such that the crack contour and its exterior are mapped onto a unit circle and its exterior respectively. Let such a mapping be denoted by $z=\omega(\zeta)$. Denoting the tips of such a crack by z_ℓ and their images in the ζ-plane by ζ_ℓ, such mapping requires $\omega'(\zeta_\ell)=0$ [5]. Hence we have the following expansion near z_ℓ

$$z-z_\ell = \omega(\zeta) - \omega(\zeta_\ell) = \frac{1}{2}(\zeta-\zeta_\ell)^2\omega''(\zeta_\ell) + \tag{6}$$

$$+ \frac{1}{6}(\zeta-\zeta_\ell)^3\omega'''(\zeta_\ell) + \cdots$$

Using the near field Westergaard solution of the tips of the crack we have, using the usual notations, with $R \to 0$

$$(7) \qquad \lim_{R\to 0} (\sigma_x+\sigma_y) = \mathrm{Re}\{K\sqrt{2/\pi R}\, e^{-i\theta/2}\}, \quad K = K_I - i\, K_{II}$$

Let the inclination of the crack tip to the x-axis be δ (Fig. 1), we have from (6)

$$(8) \qquad R\, e^{i(\theta+\delta)} = z - z_\ell = \frac{1}{2}(\zeta-\zeta_\ell)^2 \omega''(\zeta_\ell) + O(\zeta-\zeta_\ell)^3$$

and

$$(9) \qquad \omega'(\zeta) = (\zeta-\zeta_\ell)\omega''(\zeta_\ell) + O(\zeta-\zeta_\ell)^2$$

From (1) with $\phi_1'(z) = \phi'(\zeta)/\omega'(\zeta)$, $\psi_1'(z) = \psi'(\zeta)/\omega'(\zeta)$ etc. and using (8), (9) we have

$$(10) \qquad \lim_{\zeta\to\zeta_\ell} (\sigma_x+\sigma_y) = 4\, \mathrm{Re}\{\phi'(\zeta_\ell)[2R\, \omega''(\zeta_\ell) e^{i\delta}]^{-1/2}\, e^{-i\theta/2}\}$$

Comparing (7) and (10) for all θ, we have

$$(11) \qquad K = K_I - iK_{II} = 2\, \pi^{1/2}\, \phi'(\zeta_\ell)/\sqrt{e^{i\delta}\omega''(\zeta_\ell)}$$

After some manipulation using $F'(z) = f'(\zeta)/\omega'(\zeta)$ we get the corresponding expression for the anti-plane problem:

$$(12) \qquad K_{III} = i\mu\, f'(\zeta_\ell)\pi^{1/2}/\sqrt{e^{-i\delta}\omega''(\zeta_\ell)}$$

These expressions may readily be available in literature. It will be seen that δ plays a crucial role in path independent integrals. Following Irwin's approach we obtain the energy release rates:

$$(13) \qquad \mathcal{G} = \frac{4\pi}{E} \frac{\phi'(\zeta_\ell)\, \overline{\phi'(\zeta_\ell)}}{(\omega''(\zeta_\ell)\, \overline{\omega''(\zeta_\ell)})^{1/2}} \quad \text{(for plane stress)}$$

$$(14) \qquad \mathcal{G} = -\frac{\mu\pi}{2}\,\frac{f'^2(\zeta_\ell)}{e^{-i\delta}\omega''(\zeta_\ell)} \qquad \text{(for anti-plane)}$$

3. COMPUTATION OF $\mathcal{G}$ USING PATH INDEPENDENT INTEGRALS. Using path independent integrals the following expressions have been derived by Budiansky and Rice [7], for an open contour:

$$(15) \qquad J_1+iJ_2 = -\frac{2i}{E}\Big\{\int_A^B \phi_1'^2(z)dz - 2\overline{\int_A^B \phi_1'(z)\psi_1'(z)dz} - \Big[z\,\overline{\phi_1'(z)^2}\Big]_A^B\Big\}$$

$$(16) \qquad J_1 - iJ_2 = \frac{i}{2}\mu\int_A^B (F'(z))^2dz$$

for plane and anti-plane problems respectively (the term in square bracket is the jump term). Since the crack tip is inclined at an angle δ to the x-axis, we need the above integrals for an inclined coordinate system. Rotating the coordinate axes, counterclockwise through an angle δ and denoting the transformed quantities with superscript (1), we have for the plane problem [6]

$$\begin{aligned} z &= z^{(1)}e^{i\delta} \\ \Phi^{(1)}(z^{(1)}) &= \Phi(z^{(1)}e^{i\delta}) \\ (17) \qquad \Psi^{(1)}(z^{(1)}) &= e^{2i\delta}\Psi(z^{(1)}e^{i\delta}) \end{aligned}$$

where $\Phi(z)=\phi_1'(z)$, $\Psi(z)=\psi_1'(z)$ etc. Using (17) in (15) we have

$$(18) \qquad J_1^{(1)} + i\,J_2^{(1)} = e^{-i\delta}(J_1 + i\,J_2)$$

Similarly for the anti-plane problem we have

$$(19) \qquad F^{(1)'}(z^{(1)}) = e^{i\delta}\,F'(z^{(1)}e^{i\delta})$$

and (19) into (16) gives

$$(20) \qquad J_1^{(1)} - i\, J_2^{(1)} = e^{i\delta}(J_1 - i\, J_2)$$

It is easy to show that the above integrals remain invariant under translation of coordinate systems and hence without loss of generality we make an additional assumption that one of the crack tips is located at the origin of the z-plane, i.e., $z_\ell=\omega(\zeta_\ell) = 0$. Using the mapping function described previously equations (15) and (16) become

$$(21) \qquad J_1 + i\, J_2 = -\frac{2i}{E}\{\{\int_a^b \frac{\phi'^2(\zeta)}{\omega'(\zeta)}d\zeta - 2\int_a^b \frac{\overline{\phi'(\zeta)\psi'(\zeta)}}{\omega'(\zeta)}d\zeta - \left[\omega(\zeta)\left(\frac{\overline{\phi'(\zeta)}}{\omega'(\zeta)}\right)^2\right]_a^b\}$$

and

$$(22) \qquad J_1 - i\, J_2 = \frac{i}{2}\mu\int_a^b \frac{f'^2(\zeta)}{\omega'(\zeta)}\,d\zeta$$

where a, b in the ζ-plane corresponds to A and B on the z-plane, in the limit as $r\to 0$ (see Fig. 1). Now making use of (6) with $z_\ell=\omega(\zeta_\ell)=0$, (location of the crack tip at the origin of the z-plane) we see that $\psi'(\zeta_\ell)$ is finite (see eq. (31)). Equations (21), (22) can be integrated using a circular contour of vanishing radius around ζ_ℓ, (Fig. 1)

$$(23) \qquad J_1 + i\, J_2 = \frac{2\pi}{E}\left[\frac{\phi'^2(\zeta_\ell)}{\omega''(\zeta_\ell)} + 2\,\overline{\left(\frac{\phi'(\zeta_\ell)\psi'(\zeta_\ell)}{\omega''(\zeta_\ell)}\right)}\right] \quad \text{(plane)}$$

Note that the jump term in equation (21) vanishes. The corresponding result for anti-plane problem becomes:

$$(24) \qquad J_1 - i\, J_2 = -\frac{\pi}{2}\mu\, f'^2(\zeta_\ell)/\omega''(\zeta_\ell)$$

Rice has shown [2] that the energy release rate along the plane of the crack is given by $J_1^{(1)}$. Hence we have

$$(25)\qquad \mathcal{G} = J_1^{(1)} = \frac{2\pi}{E}\,\mathrm{Re}\left[e^{-i\delta}\left(\frac{\phi'^2(\zeta_\ell)}{\omega''(\zeta_\ell)} + 2\,\overline{\frac{\phi'(\zeta_\ell)\,\psi'(\zeta_\ell)}{\omega''(\zeta_\ell)}}\right)\right]$$

$$(26)\qquad \mathcal{G} = J_1^{(1)} = -\frac{\pi}{4}\,\mu\left(\frac{f'^2(\zeta_\ell)}{e^{-i\delta}\,\omega''(\zeta_\ell)} + \frac{\overline{f'^2(\zeta_\ell)}}{e^{i\delta}\,\overline{\omega''(\zeta_\ell)}}\right)$$

for the plane and anti-plane problems respectively. In the sequel it will be proved that (25) and (26) reduce to (13) and (14) respectively.

4. EQUIVALENCE OF TWO DISTINCT EXPRESSIONS FOR $\mathcal{G}$. Without loss of generality a traction free boundary condition is assumed on the crack. Since $z_\ell=\omega(\zeta_\ell)=0$ we have for ζ near ζ_ℓ

$$(27)\qquad \omega(\zeta) = \frac{1}{2}(\zeta-\zeta_\ell)^2\omega''(\zeta_\ell) + O(\zeta-\zeta_\ell)^3$$

$$\omega'(\zeta) = (\zeta-\zeta_\ell)\omega''(\zeta_\ell) + O(\zeta-\zeta_\ell)^2$$

From (1) and (2) and using polar coordinates [6], the boundary condition on the unit circle $\sigma=e^{i\theta}$ in the ζ plane becomes, (with $\Phi(\sigma)=\phi'(\sigma)/\omega'(\sigma)$)

$$(28)\qquad \sigma_r - i\tau_{r\theta} = \frac{\phi'(\sigma)}{\omega'(\sigma)} + \overline{\frac{\phi'(\sigma)}{\omega'(\sigma)}} - \overline{\frac{\sigma^2}{\omega'(\sigma)}}\,\{\overline{\omega(\sigma)}\;\Phi'(\sigma)+\psi'(\sigma)\} = 0$$

Multiplying by $\omega'(\sigma)$, which has a zero at $\sigma=\zeta_\ell$, we have,

$$(29)\qquad \phi'(\sigma) + \overline{\frac{\omega'(\sigma)}{\omega'(\sigma)}}\,\overline{\Phi'(\sigma)} - \sigma^2\,\overline{\frac{\omega(\sigma)}{\omega'(\sigma)}}\,\Phi''(\sigma) +$$

$$+\,\sigma^2\,\overline{\frac{\omega(\sigma)}{\omega'(\sigma)}}\,\frac{\omega''(\sigma)}{\omega'(\sigma)}\,\Phi'(\sigma) - \sigma^2\,\overline{\frac{\omega'(\sigma)}{\omega'(\sigma)}}\,\psi'(\sigma) = 0$$

Substituting (27) in (29) and approaching the limit $\zeta \to \zeta_\ell$ we have, (since σ and ζ_ℓ are both on the unit circle):

$$(30) \qquad \frac{1}{2}\phi'(\zeta_\ell) - \zeta_\ell^2 \frac{\omega''(\zeta_\ell)}{\overline{\omega''(\zeta_\ell)}} \overline{\phi'(\zeta_\ell)} + \zeta_\ell^4 \frac{\omega''(\zeta_\ell)}{\overline{\omega''(\zeta_\ell)}} \psi'(\zeta_\ell) = 0$$

Solving for $\psi'(\zeta_\ell)$ we have

$$(31) \qquad \psi'(\zeta_\ell) = \frac{1}{\zeta_\ell^2} \overline{\phi'(\zeta_\ell)} - \frac{1}{2\,\zeta_\ell^4} \frac{\overline{\omega''(\zeta_\ell)}}{\omega''(\zeta_\ell)} \phi'(\zeta_\ell)$$

Since ζ_ℓ is on the unit circle we write $\zeta_\ell = e^{i\alpha_\ell}$. From (8) it is clear that the top and bottom lines of the crack form a tangent to the unit circle at ζ_ℓ in the ζ-plane. Consider a point on the bottom line close to the crack tip, point A in Fig. 1, $\theta = -\pi$, $\beta = \alpha_\ell - \frac{\pi}{2}$ (as $r \to 0$), equation (8) yields

$$(32) \qquad R\, e^{i\delta} = \frac{1}{2} r^2 \zeta_\ell^2 \omega''(\zeta_\ell)$$

Taking complex conjugate of (32) and eliminating R and r we have

$$(33) \qquad e^{i\delta} = \left(\omega''(\zeta_\ell)/\overline{\omega''(\zeta_\ell)}\right)^{1/2} \zeta_\ell^2$$

Substituting (31) and (33) into (25) and rearranging we get

$$J_1^{(1)} = \mathcal{G} = \frac{2\pi}{E} \mathrm{Re}\left[\frac{2\,\phi'(\zeta_\ell)\,\overline{\phi'(\zeta_\ell)}}{\left(\omega'(\zeta_\ell)\,\overline{\omega''(\zeta_\ell)}\right)^{1/2}} + \right.$$

$$\left. + \left(\omega''(\zeta_\ell)\overline{\omega''(\zeta_\ell)}\right)^{1/2} \left(\frac{\phi'^2(\zeta_\ell)}{\zeta_\ell^2 \omega''^2(\zeta_\ell)} - \frac{\overline{\phi'^2(\zeta_\ell)}}{\overline{\zeta_\ell^2}\,\overline{\omega''^2(\zeta_\ell)}} \right) \right]$$

In which the first term is real and the second term is imaginary, hence

$$(34) \qquad J_1^{(1)} = \mathcal{G} = \frac{4\pi}{E} \frac{\phi'(\zeta_\ell)\ \overline{\phi'(\zeta_\ell)}}{[\omega''(\zeta_\ell)\ \overline{\omega''(\zeta_\ell)}]^{1/2}}$$

which is the same as equation (13) obtained by Irwin's approach. Equation (31) indicates that $\psi'(\zeta_\ell)$ is finite. This primarily is due to our assumption that the crack tip is at the origin on z-plane i.e., $z_\ell=\omega(\zeta_\ell)=0$. (See appendix for a general case).

For the antiplane problem, using cylindrical coordinates, we have

$$(35) \qquad \tau_{rz} - i\ \tau_{\theta z} = \mu\ F'(z)e^{i\theta}$$

Again assuming a traction free boundary condition in the vicinity of the crack tip we have

$$\mathrm{Im}\{\mu\ F'(z)e^{i\delta}\} = \frac{\mu}{2i}\{e^{i\delta}F'(z) - e^{-i\delta}\ \overline{F'(z)}\} = 0$$

and using the mapping function (i.e., $F'(z)=f'(\zeta)/\omega'(\zeta)$) we have

$$(36) \qquad e^{i\delta}\ f'(\zeta)/\omega'(\zeta) = e^{-i\delta}\ \overline{f'(\zeta)}/\overline{\omega'(\zeta)}, \qquad \zeta=\sigma=e^{i\theta}$$

In the limit as $\zeta\to\zeta_\ell$, using (27) and (33), (36) becomes

$$(37) \qquad f'(\zeta_\ell)\zeta_\ell^2 = -\ \overline{f'(\zeta_\ell)}$$

Using (37) and (33) into (26) we get

$$(38) \qquad \mathcal{G} = J_1^{(1)} = -\frac{\mu\pi}{2} \frac{f'^2(\zeta_\ell)}{e^{-i\delta}\ \omega''(\zeta_\ell)}$$

which is the same as equation (14) obtained by Irwin's approach.

APPENDIX

In the text we have located one of the crack tips at the origin of the z-plane. This leads to finiteness of $\psi'(\zeta_\ell)$ for the plane problem. For an arbitrary location, $\psi'(\zeta_\ell)$ involve singularity and the integrals together with the jump term become singular as we take the contour in the ζ-plane of vanishing radius. Here we show the annihilation of these terms.

Using Taylor's expansion around the crack tip ζ_ℓ with $\omega'(\zeta_\ell)=0$ we have (referring to equation (27))

$$\omega(\zeta) = \omega(\zeta_\ell) + \frac{1}{2}(\zeta-\zeta_\ell)^2\omega''(\zeta_\ell) + \frac{1}{6}(\zeta-\zeta_\ell)^3\omega'''(\zeta_\ell) + \cdots$$

$$\text{(A-1)} \qquad \frac{1}{\omega'(\zeta)} = \frac{1}{\omega''(\zeta_\ell)(\zeta-\zeta_\ell)}\{1 + c_1(\zeta-\zeta_\ell) + c_2(\zeta-\zeta_\ell)^2 + \cdots\}$$

$$\left(\frac{1}{\omega'(\zeta)}\right)^2 = \frac{1}{\omega''^2(\zeta_\ell)(\zeta-\zeta_\ell)^2}\{1 + 2\,c_1(\zeta-\zeta_\ell) + c_2'(\zeta-\zeta_\ell)^2 + \cdots\}$$

where $c_1 = -(1/2)\omega'''(\zeta_\ell)/\omega''(\zeta_\ell)$ etc.

Contour integrations and the jump terms are given by (Fig. 1)

$$\int_a^b \frac{d\zeta}{\zeta-\zeta_\ell} = \pi i, \quad \int_a^b \frac{d\zeta}{(\zeta-\zeta_\ell)^2} = -\frac{2\zeta_\ell}{r}, \quad \int_a^b \frac{d\zeta}{(\zeta-\zeta_\ell)^3} = 0$$

$$\text{(A-2)}$$

$$\left(\frac{1}{\zeta-\zeta_\ell}\right)_a^b = \frac{2\zeta_\ell}{r}, \quad \left(\frac{1}{(\zeta-\zeta_\ell)^2}\right)_a^b = 0, \quad (\zeta-\zeta_\ell) = re^{i\beta}$$

From boundary condition (29) we have ($\sigma=e^{i\theta}$ is on the unit circle)

$$\text{(A-3)} \qquad \psi'(\sigma) = \frac{1}{\sigma^2}\frac{\overline{\omega'(\sigma)}}{\omega'(\sigma)}\phi'(\sigma) + \frac{1}{\sigma^2}\overline{\phi'(\sigma)} - \frac{\overline{\omega(\sigma)}}{\omega'(\sigma)}\phi''(\sigma) + \frac{\overline{\omega(\sigma)}\,\omega''(\sigma)}{\omega'^2(\sigma)}\phi'(\sigma)$$

Substituting (A-1) in (A-3) and using analytic continuation we have

$$\lim_{\zeta\to\zeta_\ell} \psi'(\zeta) = \frac{1}{(\zeta-\zeta_\ell)^2}\ \frac{\overline{\omega(\zeta_\ell)}}{\omega''(\zeta_\ell)}\ \phi'(\zeta_\ell) + \psi'^{*}(\zeta) \tag{A-4}$$

where $\psi'^{*}(\zeta)$ is regular at ζ_ℓ. Hence

$$\lim_{\zeta\to\zeta_\ell} \frac{2\phi'(\zeta)\psi'(\zeta)}{\omega'(\zeta)} = \frac{2\overline{\omega(\zeta_\ell)}}{\omega''^{2}(\zeta_\ell)}\ \frac{\phi'(\zeta_\ell)}{(\zeta-\zeta_\ell)^3} \tag{A-5}$$

$$\{\phi'(\zeta_\ell) + [\phi''(\zeta_\ell) + c_1\phi'(\zeta_\ell)](\zeta-\zeta_\ell) + 0\ (\zeta-\zeta_\ell)^2\}$$

As can be seen from (A-5) and (A-2), (21) has singular terms as $r\to 0$ in the second integral given by

$$-\frac{2\zeta_\ell}{r}\ \frac{2\ \overline{\omega(\zeta_\ell)}}{\omega''^{2}(\zeta_\ell)}\ [\phi'(\zeta_\ell)\phi''(\zeta_\ell) + c_1\phi'^{2}(\zeta_\ell)] \tag{A-6}$$

Considering now the complex conjugate of the jump term in (21), we have

$$\overline{\omega(\sigma)}\frac{\phi'^{2}(\sigma)}{\omega'^{2}(\sigma)} = 0\ \left(\frac{1}{\zeta-\zeta_\ell}\right)^2 + \frac{2\overline{\omega(\zeta_\ell)}}{\omega''^{2}(\zeta_\ell)}\ [\phi'(\zeta_\ell)\phi''(\zeta_\ell) + c_1\phi'^{2}(\zeta_\ell)]\ \frac{1}{\zeta-\zeta_\ell} + \cdots \tag{A-7}$$

From (A-7) and (A-2) we have the jump term from (21),

$$\lim_{r\to 0}\left[\overline{\omega(\zeta)}\frac{\phi'^{2}(\zeta)}{\omega'^{2}(\zeta)}\right]_a^b = \frac{2\zeta_\ell}{r}\ \frac{2\ \overline{\omega(\zeta_\ell)}}{\omega''^{2}(\zeta_\ell)}\ [\phi'(\zeta_\ell)\phi''(\zeta_\ell) + c_1\phi'^{2}(\zeta_\ell)] \tag{A-8}$$

which is the negative of (A-6) and hence leads to annihilation of the singular terms in (21).

REFERENCES

1. Irwin, G. R., "Analysis of Stresses and Strains Near the End of a Crack Traversing a Plate," Journal of Applied Mechanics, Vol. 24, 1957, pp. 361-364.

2. Rice, J. R., "Mathematical Analysis in the Mechanics of Fracture," Chapter 3 'Fracture, An Advanced Treatise,' Edited by H. Liebowitz, Vol. II, 1968, Academic Press.

3. Bueckner, H. F., "The Propagation of Cracks and the Energy of Elastic Deformation," Transactions, ASME, Vol. 80, 1959, pp. 1225-1230.

4. Sanders, J. L., "On the Griffith-Irwin Fracture Theory," Journal of Applied Mechanics, Vol. 27, No. 2, June 1960.

5. Hussain, M. A., Pu, S. L., Underwood, J., "Strain Energy Release Rate for Crack Under Combined Mode I and Mode II," Fracture Analysis, ASTM STP 560, American Society for Testing and Materials, 1974, pp. 2-28.

6. Muskhelishvili, N., "Some Basic Problems of the Mathematical Theory of Elasticity," Noordhoff, Groningen, 1963.

7. Budiansky, B., and Rice, J. R., "Conservation Laws and Energy Release Rates," Journal of Applied Mechanics, Vol. 40, No. 1, March 1973.

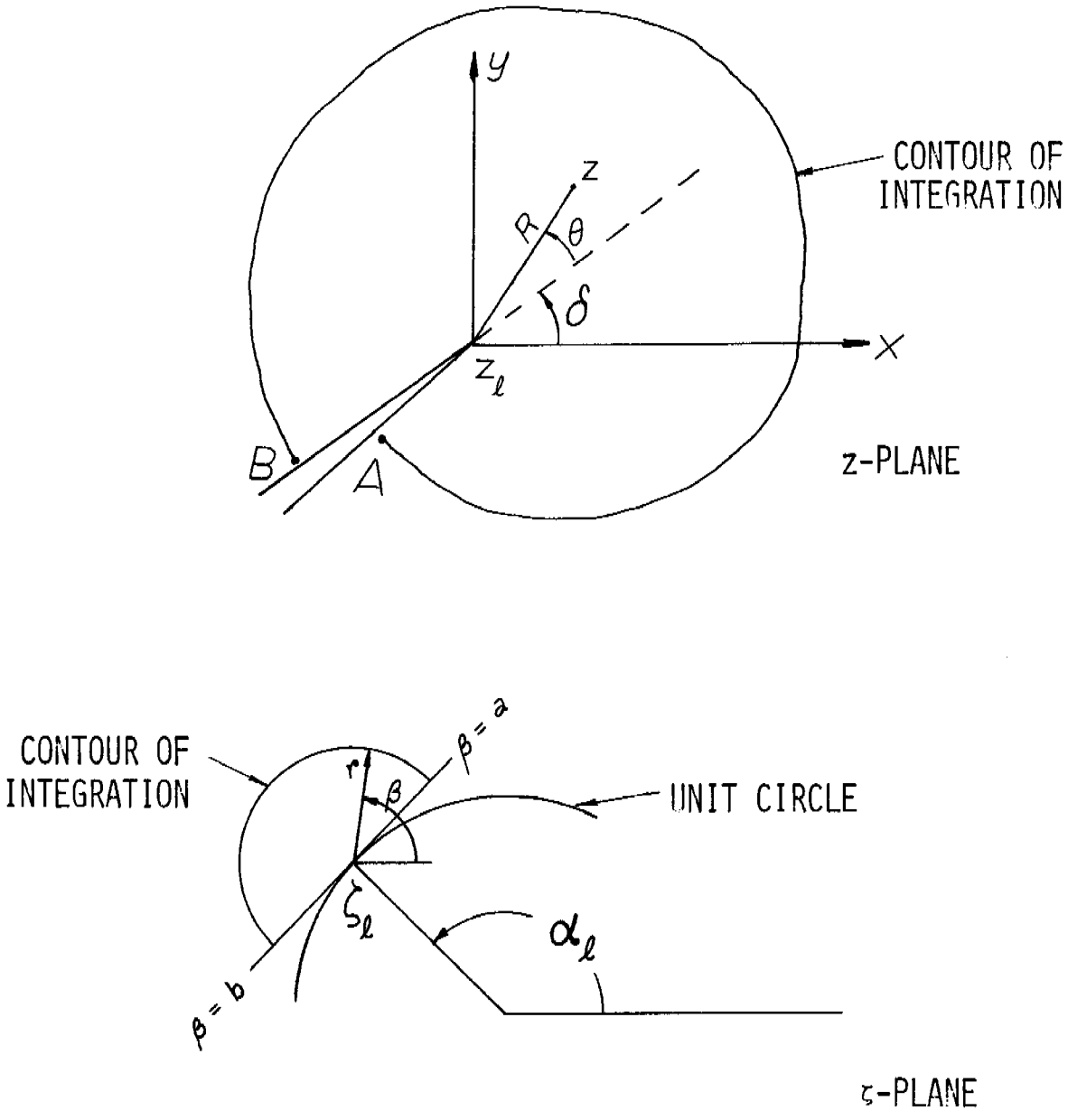


Figure 1. Mapping and contours of integration. In the limit as the contours shrink, $\beta=a$ and $\beta=b$ approach the images of A and B on the ζ-plane.

DEVELOPMENT AND APPLICATION OF DYNAMIC MATHEMATICAL MODELS FOR EVALUATION OF MILITARY SYSTEMS, FORCES AND DOCTRINE

Roger F. Willis

US Army Combined Arms Combat Developments Activity
Fort Leavenworth, Kansas

ABSTRACT. This paper starts with a review of alternative combined arms study objectives for which flexible mathematical models are required and definition of the characteristics that such models should possess. Specific mathematical models, involving sets of simultaneous differential equations, and their solutions (as well as indications of how these models can be applied to real problems) are presented in the following areas:

1. tank combat, including artillery
2. anti-tank weapon employment
3. air defense suppression
4. helicopter-air defense duels
5. support weapon employment rates

1. This paper will cover a variety of classes of mathematical models of military operations. Before doing that we will briefly discuss the types of studies we do in TRADOC and explain how mathematical models fit into our study methodologies.

2. To meet combat development objectives we carry out many different types of studies, some using large, expensive war games or computer simulations. These types of applications are listed in Table 1. Examples of specific studies in some of these categories are:

Type 2. Comparison of a US division supported by scatterable mines with a US division without scatterable mines.

Type 4. Comparison of the new XM-1 tank with the standard M-60 tank.

Type 5. Evaluate alternative tactics and procedures for air defense suppression.

Type 6. Determine the optimum mix of intelligence-gathering systems.

Type 7. Determine range requirements for artillery.

Type 8. Determine the numbers of mortars of each caliber required in each type of maneuver battalion.

Type 9. Develop effective countermeasures against antitank missiles.

Type 10. Determine the impact of political constraints on tactical nuclear doctrine.

3. First we consider the specific example of a type 2 study -- a US division without scatterable mines (Force A) compared with the same US division with scatterable mines (Force B). Suppose we use a large, complex, computer-assisted war game, play two games, and get the results summarized at Figure 1. We played about a day of combat with Force A and, starting over against the same enemy force, played about a day of combat with Force B. Force B appears to be more effective based on the first three measures of effectiveness, but Force A is more effective if we use any one of the last three measures. Thus, as we would expect, the study results are sensitive to the measure of effectiveness used. A second point is that the differences measured between Force A and Force B might not be large enough to be significant (e.g., 27 hours compared with 23 hours).

4. These aggregated results are backed up by a great variety of detail that can be studied to develop insights concerning the real differences in capabilities between Force A and Force B. In particular both games are based on hundreds of assumptions, some of which might have major impacts on the combat results. This raises the question: How sensitive are the results to variations in the assumptions? Some of the assumptions are key elements in the scenario.

5. For every study we are required to use a standard scenario. This could give us results as illustrated in Figure 2, in which Force A and Force B are not significantly different against the standard size of enemy attack. The measure of effectiveness used might be, for example, the percent of US tanks surviving at the end of 3 days. If we use a larger enemy force and re-run both the Force A game (45%) and the Force B game (65%) then Force B becomes significantly better than Force A. Another scenario factor is the amount of close air support available to the US force. The lower half of Figure 2 shows that increasing the amount of close air support above standard still leaves Force A about equal to Force B in effectiveness. However, reducing close air support makes Force A significantly more effective than Force B. Thus this scenario factor reverses the trend.

6. In order to play these six extra games, for investigating the sensitivity of study results to only two factors, might take 15 people

3 or 4 months. It is clear that we need more efficient ways of doing sensitivity analysis. In particular, we need a variety of flexible mathematical models in many different subject areas, as follows:

a. ground combat

b. artillery

c. attack helicopters

d. close air support

e. battlefield interdiction

f. tactical nuclear

g. air defense

h. reserves

i. replacements

j. resupply

k. air reconnaissance

l. ground surveillance

m. information flow

n. airmobile operations

o. engineer operations

p. transportation

q. command and control

7. These models are characterized not by the nature of the mathematical formulation but by the categories of military operations or systems that the models represent. Such models are needed to provide analytical support for different types of studies, in a variety of ways to be discussed later.

8. The first example of a model presented here involves the first two categories, ground combat (primarily direct fire) and artillery. The four differential equations describing this model appear in Figure 3. The basic assumptions are:

a. Tanks fire at tanks and artillery is employed against tank units and also in counterbattery fire against opposing artillery.

b. The rate at which each side loses tanks is proportional to the number of tanks employed by the other side (and still surviving) and proportional to the number of artillery tubes the other side is using against tanks.

c. The rate at which each side loses artillery is proportional to the number of artillery tubes the other side employs against artillery.

9. The factors in this model are defined as follows:

B(t) = number of Blue tanks surviving at time t

R(t) = number of Red tanks surviving at time t

H(t) = number of Blue artillery tubes surviving at time t

G(t) = number of Red artillery tubes surviving at time t

J = rate at which each Red tank kills Blue tanks

K = rate at which each Blue tank kills Red tanks

L = expected number of Blue tanks killed per Red artillery round fired.

M = expected number of Red tanks killed per Blue artillery round fired

r = rate of Red artillery fire

p = rate of Blue artillery fire

E = expected number of Blue artillery tubes killed per Red artillery round fired

F = expected number of Red artillery tubes killed per Blue artillery round fired

f = fraction of surviving Red artillery tubes employed against tanks

g = fraction of surviving Blue artillery tubes employed against tanks.

Note that the last two factors (f and g) are tactical decision parameters, in contrast with the other variables that represent numbers available and performance characteristics of various weapon systems. Therefore this model does represent one aspect of tactics explicitly.

10. The solutions of this model are in the following form:

a. Number of Blue tanks surviving versus time

b. Number of Red tanks surviving versus time

c. number of Blue artillery tubes surviving versus time

d. number of Red artillery tubes surviving versus time

These solutions have been obtained analytically, without using a computer.

11. This model has a parallel in any war game in that periodically the Blue and Red commanders must decide (unless it is programmed to happen automatically) what part of their surviving artillery force to employ in counterbattery fire. To some extent this decision would be influenced by target acquisition, but this could also be included in the model in Figure 3 by expanding the aggregated coefficients J, K, L, M, E, and F to include several types of target acquisition capability. One difference is that in a war game the players have no way to find optimum tactics whereas this model could be used to find optimum values for f (Red tactic) and for g (Blue tactic). The values of f and g could also be varied from time period to time period. Note that the vulnerability of Red tanks to Blue artillery fire is included in the factor M and the vulnerability of Red artillery to Blue artillery fire is included in the factor F.

12. The advantage in having these solutions as explicit functions of all the assumptions and parameters is that one can carry out a wide variety of analytical variations before assuming numerical values for any of the factors. Some of these variations are:

a. Values versus time

(1) number surviving

(2) percent surviving

(3) force ratio of survivors

(4) number lost

(5) percent lost

(6) loss ratio

b. Time required for mission accomplishment, e.g., Red losses = 30 percent.

c. Sensitivity of "values" or "times" to force sizes, tactics, or other assumptions.

d. Optimize mix of systems

e. Optimize tactics

f. Catalog situations giving the same results

In the case of tanks the six "values" listed under a. above are as follows:

a. Number surviving: $B(t)$

b. Percent surviving: $100\,\frac{B(t)}{B_0}$

c. Force ratio: $R(t)/B(t)$

d. Number lost: $B_0 - B(t)$

e. Percent lost: $100\left(\frac{B_0 - B(t)}{B_0}\right)$

f. Loss ratio: $L = \frac{R_0 - R(t)}{B_0 - B(t)}$

B_0 and R_0 are the numbers of tanks available at time o. The time required to kill at least 30 percent of the Red tanks can be found by solving the following equation for time t:

$$B(t) = 0.70\ B_0$$

The sensitivity of the loss ratio to initial force sizes can be determined from

$$\frac{dL}{dR_0} \quad \text{and} \quad \frac{dL}{dB_0}.$$

13. In this model, tactics can be optimized for Blue by finding that value of g that will accomplish one of the following goals:

a. Maximize B(t) at a given t (e.g., 2 hours)

b. Minimize R(t)/B(t)

c. Maximize $L = \frac{R_0 - R(t)}{B_0 - B(t)}$

d. Minimize time required to kill at least X percent of the Red tanks.

e. Maximize some weighted average of Blue tanks surviving and Blue artillery surviving at a given time t. (e.g., maximize a B(t) + bH(t)).

14. Using the explicit solutions of this model one can catalog all situations (combinations of values of input factors) that give the same results, such as 20 percent of the Blue tanks lost by 4 hours after the battle starts. One example would be all combinations of Blue artillery effectiveness against tanks (M) and initial number of Red tanks (R_0) that result in no more than 5 percent of Blue tanks lost at 4 hours.

15. The above paragraphs give specific examples of the types of calculations that can be developed from the type of model illustrated. In more general terms eight uses of mathematical models of military operations are:

a. Compare alternative systems

b. Compare alternative tactics

c. Optimize tactics

d. Tradeoffs between system parameters

e. Narrow down number of alternatives to be gamed

f. Select ranges of values for inputs

g. Select likely enemy tactics

h. Sensitivity analysis

Both alternative systems and alternative tactics can be compared in a force-on-force context. Optimizing tactics has been discussed in the earlier example. The fourth type of use could be illustrated by tradeoffs between lethality and rate of fire. Type e might be narrowing down the number of levels of close air support to use in actual gaming -- levels that are feasible and also have been shown by mathematical modelling to have significant impact on the particular types of study results of interest. Type f could involve values for an enemy weapon performance parameter for which we do not have good estimates. Type g could be, for example, weapon-target priorities.

16. For discussing specific models in this paper we have defined classes of combat models according to the eight factors in Table 2. Factor 3 refers to variation with time (or with weapon range, for example). We have given several examples earlier of aggregating or expanding coefficients to represent specific elements such as vulnerability, rate of fire, target acquisition, etc. (Factor 4). Factor 5 concerns allocations of weapons against targets and the replacements in Factor 6 are either personnel or equipment.

17. Setting factors 1 - 6 at two levels each we have 64 classes of models. Thus, for example, considering only tank and antitank direct fire weapons would give 64 classes. However, factors 7 and 8 raise the number of classes of models to well above 256. For example in factor 8 we could have 1, 2, 3, or 4 types of Blue systems and 1, 2, 3, or 4 types Red systems in the models. We have developed at least 50 models in various classes with the taxonomy of Table 2. For factors 7 and 8 the types of weapons of major interest are listed in Table 3.

18. Now we present a model designed to study tactical interactions between tanks and antitank missiles, in Figure 4. The situation is as follows: Blue has both tanks and antitank missiles; Red has only tanks; how should Red employ its tanks? The fraction of Red tank fire directed against Blue tanks is f. What is the optimum value of f?

19. By solving these three equations to obtain the remaining weapons on both sides as explicit functions of time and of all the other parameters (including the initial strengths) we can find out how the optimum values of f depend on all these quantities and on time (duration of combat). The solutions are given in Figures 5 and 6. In order to optimize f, Red could try to maximize the number of Red tanks surviving or the force ratio surviving;or to minimize the number of Blue tanks surviving, the number of Blue antitank missiles surviving or some linear combination of these Blue weapons.

20. In Table 4 we list the major interactions in the air defense suppression model presented in Figure 7. Both sides carry out all six operations listed in Table 4 and thus Red and Blue each have two allocation problems:

a. Allocate artillery fire against air defense weapons or in counterbattery fire.

b. Allocate aircraft (close air support) against tanks or against air defense weapons.

Thus complete analysis of results from this model would involve a game matrix in which each side has at least 9 alternatives. Artillery versus air defense at high, medium or low values times aircraft versus air defense at high, medium or low values.

21. The allocation factors in Figure 7 are defined as follows:

f_R = fraction of Red aircraft used against Blue tanks

f_B = fraction of Blue aircraft used against Red tanks

g_R = fraction of Red artillery used against Blue air defense

g_B = fraction of Blue artillery used against Red air defense.

The kill factors H, J, C and D in the fifth and sixth equations could be expanded to include target acquisition.

22. This model has also been solved analytically. A typical battle is illustrated in Figure 8, based on the following assumptions:

F = .005	G = .010	f_R = .50
R = .010	B = .010	f_B = .50

$K = .010$ $g_R = .50$ $1 - g_R = .50$

$E = .015$ $g_B = .05$ $1 - g_B = .95$

$H = .02$ $J = .03$ $1 - f_R = .50$

$C = .01$ $D = .03$ $1 - f_B = .50$

$$W(Y_7) = -.04Y_7$$

$$Z(X_7) = -.04X_7$$

Figure 8 shows the Blue and Red survivors of each type of weapon versus time. After 30 minutes the Blue tanks are down to less than 50% of initial strength, Red has lost 8 artillery tubes and Blue has almost no air defense left. The Red tactic was to use half of its artillery and half of its aircraft against Blue air defense.

23. In the above example it was assumed that aircraft losses are directly proportional to the density of air defenses. Some insights concerning the conditions under which this might be a valid assumption can be obtained from duel models and force-on-force models of attack aircraft versus complexes of air defense systems. The next model is a stochastic duel between an attack helicopter and a single air defense weapon, including the following factors:

a. detection capability

b. warning of being detected

c. rate of fire

d. weapon accuracy

e. weapon lethality

f. vulnerability

The capability of the air defense weapon to detect the helicopter and the capability of the helicopter to detect the air defense weapon are both included, as well as the likelihood that the helicopter will be aware that it is detected (e.g., painted by radar). Rates of fire for both weapons are included as well as accuracy and lethality factors.

24. The basic inputs for the model are listed in Table 5. Using these inputs the model calculates, for the total encounter of T minutes, the probability that the air defense weapon is killed, the probability that the helicopter is hit (but not necessarily killed) and the probability that the helicopter is killed. This third measure is illustrated in Figure 9, based on the inputs defined earlier. Using this model one can study the sensitivity of helicopter (or air defense weapon) success to variations in a number of system performance and tactical parameters, such as the following:

flight tactics
bursts fired
rounds per burst

detection capability
warning equipment
jamming

readiness, decisions, response times

round lethality
burst pattern
rate of fire
dispersion.

25. The final example of a combat model illustrates how models can be developed to represent any one of three concepts:

a. changing resource employment rates

b. diminishing returns

c. increasing or decreasing effectiveness with time.

The model involves both Red and Blue employing support weapons at rates that change with time (see Figure 10). The support weapons could be attack aircraft, mortars, artillery, tactical nuclear weapons or mines. Blue and Red employment rates are each characterized by three parameters. These exponentially decreasing rates, together with the single weapon effectiveness, combine to kill the basic elements (e.g., tanks) in addition to the ability of these basic elements to kill each other, as illustrated in the upper half of Figure 11. The Blue tank survivors versus time, as shown in the lower part of Figure 11, is one half of the solution of this model. A similar expression is available for Red tank survivors.

26. The Blue force could use these expressions to find optimum values for n, m, and b (representing the employment tactics for support weapons).

This problem is a special case in the calculus of variations -- to find the optimum function from a given class of functions.

27. We have developed a facility for rapid design of mathematical models tailored to particular problems. In Tables 6 and 7 some of the design variables are reviewed. Sensitivity analysis of uncertainties is treated either as variations of key factors in deterministic models or as random variables in stochastic models. Representation of optimum tactics, in terms of allocation decisions, has been illustrated in several different models. Clearly, various tradeoffs between quantity (e.g., number of tanks initially available) and quality (in terms of either vulnerability or killing power) can be easily studied with these models -- finding those combinations of quantity and quality that yield the same results (e.g., number of Red tanks killed by the end of three hours).

28. Using cost constraints and cost-quantity relations the models can be used to find optimum mixes and break-even points -- using combat measures of effectiveness that are consistent with the other applications. A new objective will be to incorporate adaptive tactics into some of the models, so that the model can learn and develop new tactical decisions not anticipated during the formulation of the model. Comparisons of equal cost forces on the basis of differences in effectiveness would be straightforward. In addition, with these models we can set an effectiveness goal and calculate the number of systems that are required (and their total cost) for an alternative force.

29. Using appropriate members of the following families of functions:

a. $be^{-cx^k} + f$

b. bx^k

c. $b\left[1 - e^{-cx^k}\right]$

most cases of diminishing returns, increasing or decreasing effectiveness and increasing or decreasing resource commitment rates can be represented, as illustrated in the last model presented. Examples of increasing or decreasing effectiveness are: new weapons employed for the first time; increasing hit probability as range decreases; targets taking cover; voluntary suppression.

30. Expansion of coefficients was illustrated in several cases, to represent system parameters, environmental constraints and tactical alternatives. Further model developments will incorporate more interdependence between coefficients, such as artillery fire making targets less vulnerable to direct fire or mines making targets more vulnerable to direct fire. Various non-linear relations, such as that between aircraft and air defense are candidates for variation. Alternative levels of aggregation could be applied separately to the representations of direct fire, fire support, target acquisition, etc.

31. A model could include separate equations for each echelon: company, battalion, brigade. Movement of forces, depending on strengths, time, and terrain could also be included. In item 21 of Table 7, "lateral units" refers to representation of mutual support (or failure to support). Future model developments will also include more explicitly the following features:

a. mission

b. doctrine and tactics

c. Situation

(1) Physical environment

(2) enemy alternatives

d. conditional decision-making

e. system performance

f. constraints

g. target acquisition

h. supply

i. fire support

j. weather

k. communications

l. organization

m. training

32. A number of approaches for validation of the types of mathematical models presented in this paper are being considered. These are:

a. specific input data from tests

b. high resolution simulations

c. other mathematical models

d. DIVWAG submodels

e. results of other analytical studies

f. historical data

g. hand simulations

h. military judgment

Table I

Types of Applications of Models

1. Effectiveness current forces
2. Compare alternative forces
3. Effectiveness current systems
4. Compare alternative systems
5. Evaluate alternative strategies, tactics, concepts, deployments, doctrine
6. Tradeoffs between systems, optimize force mixes
7. System performance requirements
8. Force requirements; numbers of systems required
9. Analyze enemy tactics and friendly force countermeasures
10. Impact of constraints (political, environmental, operational)
11. Unit performance inputs for theater-level models
12. Define test data required and extrapolate test data

Measure of Effectiveness		US Alternatives	
		Force A	Force B
Enemy losses	Tanks	580	650
	Personnel	9,800	9,950
Time required for enemy mission (hrs)		23	27
US losses:	Tanks	250	300
	Personnel	4,200	5,800
Tank loss ratio	Enemy / US	2.32	2.17

Figure 1. Typical Divwag Results

Assumption	Effectiveness of US forces	
	Force A	Force B
Size of enemy attack		
1. Standard	67	71
2. Larger	45	65
Amount of US close air support		
1. Standard	67	71
2. Larger	74	76
3. Smaller	56	40

Figure 2. Standard scenario problem

Tank loss rates	Artillery loss rates
$\frac{dB}{dt} = JR + fLrG$	$\frac{dH}{dt} = Er\,(1-f)\,G$
$\frac{dR}{dt} = KB + gMpH$	$\frac{dG}{dt} = Fp\,(1-g)\,H$

Figure 3. Analytical model

Table 2. Classes of Two-sided Combat Models

1. Deterministic or stochastic
2. Linear or non-linear
3. Coefficients constant or variable
4. Coefficients aggregated or not
5. Allocation decisions: No or Yes
6. Types of weapons:

   Direct fire only or not
7. Replacements: No or yes
8. Number of types of systems:

   Blue

   Red

Table 3. Types of Weapons

Direct Fire	Other
Tanks	Artillery
Machine Guns	Mortars
Infantry	Helicopters
Anti-tank Missiles	Aircraft
	Air Defense
	Mines
	Nuclear

Red Tank Loss Rate

$$\frac{dX}{dt} = BY + CZ$$

Blue Tank Loss Rate

$$\frac{dY}{dt} = EfX$$

Blue Anti-tank Missile Loss Rate

$$\frac{dZ}{dt} = H(1 - f)X$$

Initial Numbers: Xo, Yo, Zo

Red Tactic: Fire fraction f at tanks.

Figure 4. Anti-tank Weapon Employment Model

Red Tanks Surviving

$$X(t) = Xo \cosh\left(t\sqrt{BEf + CH(1-f)}\right) + \frac{BYo + CZo}{\sqrt{BEf + CH(1-f)}} \sinh\left(t\sqrt{BEF + CH(1-f)}\right)$$

Blue Anti-tank Missiles Surviving

$$Z(t) = \frac{BEfZo - BH(1-f)Yo}{BEf + CH(1-f)} + \frac{CH(1-f)Zo + BH(1-f)Yo}{BEf + CH(1-f)} \cosh\left(t\sqrt{BEf + CH(1-f)}\right) + \frac{H(1-f)Xo}{\sqrt{BEf + CH(1-f)}} \sinh\left(t\sqrt{BEf + CH(1-f)}\right)$$

Figure 5. Tank and ATM Survivors

Blue Tanks Surviving

$$Y(t) = \frac{CH(1-f)Yo - ECfZo}{BEf + CH(1-f)} + \frac{EBfYo + ECfZo}{BEf + CH(1-f)} \cosh\left(t\sqrt{BEf + CH(1-f)}\right) + \frac{EfXo}{\sqrt{BEf + CH(1-f)}} \sinh\left(t\sqrt{BEf + CH(1-f)}\right)$$

Force Ratio Surviving:

$$\frac{X(t)}{a\,Y(t) + bZ(t)}$$

Figure 6. Anti-tank Weapon Employment Model

Table 4. Air Defense Suppression Model

1. Tanks	Attack	Tanks
2. Artillery	Attack	Artillery
3. Artillery	Attack	Air Defense
4. Air Defense	Attack	Aircraft
5. Aircraft	Attack	Tanks
6. Aircraft	Attack	Air Defense

TANKS	Blue	$\frac{dX_1}{dt} = - FY_1 - Gf_R Y_6$
	Red	$\frac{dY_1}{dt} = - RX_1 - Bf_B X_6$
ARTILLERY	Blue	$\frac{dX_4}{dt} = - K(1 - g_R) Y_4$
	Red	$\frac{dY_4}{dt} = - E(1-g_B) X_4$
AIR DEFENSE WEAPONS	Blue	$\frac{dX_7}{dt} = - Hg_R Y_4 - J(1-f_R) Y_6$
	Red	$\frac{dY_7}{dt} = - C_{gB} X_4 - D (1-f_B) X_6$
AIRCRAFT	Blue	$\frac{dX_6}{dt} = W (Y_7)$
	Red	$\frac{dY_6}{dt} = Z (X_7)$

Figure 7. Loss Rates
Air Defense Suppression Model

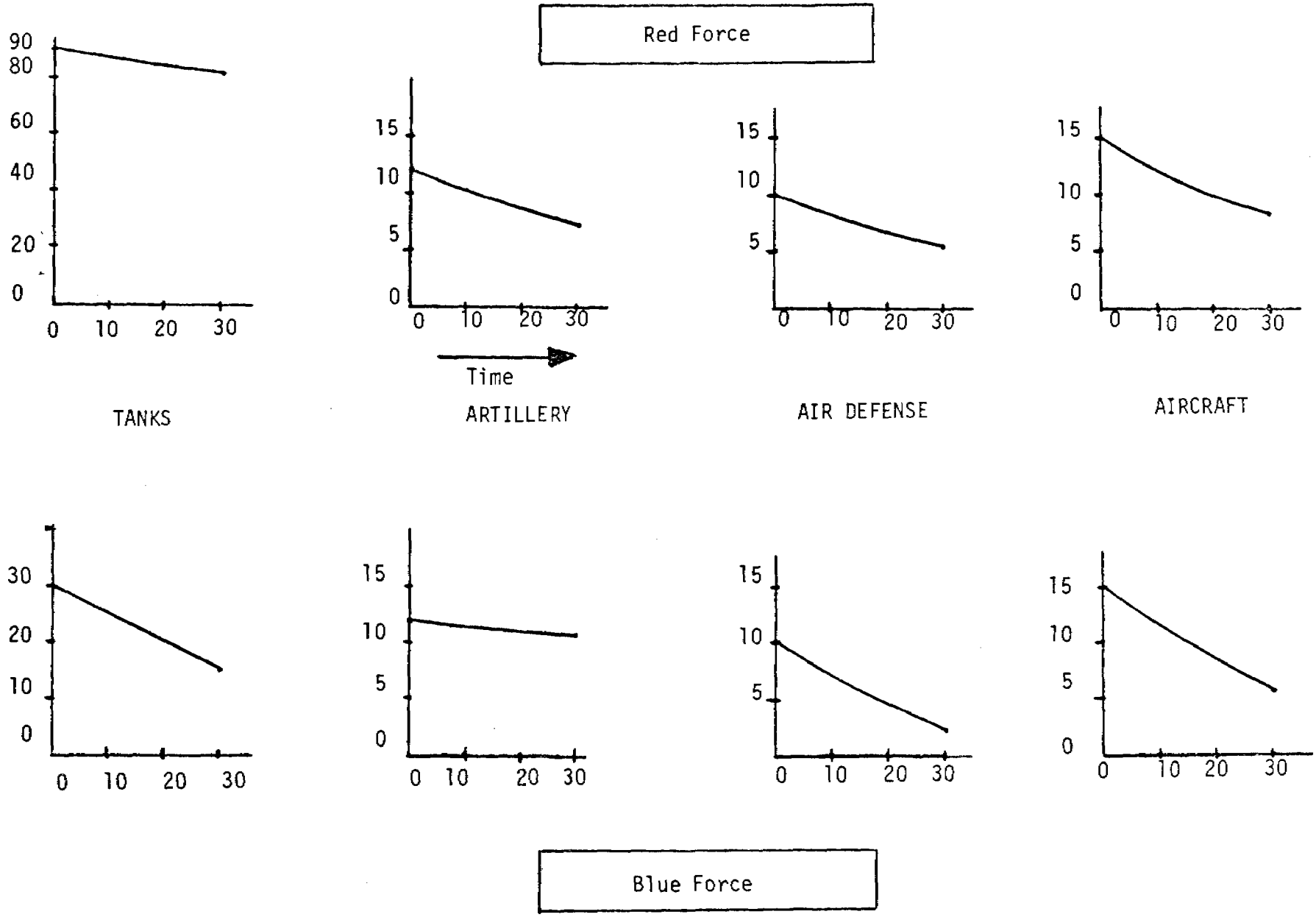


Figure 8. Air Defense Suppression Model

Table 5. Duel Model Inputs

	Helicopter	AD Weapon
Hit Probability:		
Warned target		A
Unwarned target		B
Kill Probability	C	D
Detect First Probability	E	1-E
Probability Aware of Detection	F	
Rate of Fire	V	R
Duration of Encounter	T	T

Probability Helicopter Killed:

$$DEP_1 + D(1-E)\left\{F\left[P_2 - (B-A)\right] + (1-F)\,P_2\right\}$$

Where

$$P_1 = \frac{(1-C)\left[1-(1-B)^{R/V}\right]}{1-(1-C)(1-B)^{R/V}}\left\{1-(1-B)^{TR}(1-C)^{TV}\right\}$$

$$P_2 = \frac{1-(1-B)^{R/V}}{1-(1-C)(1-B)^{R/V}}$$

Figure 9. Sample Model Output

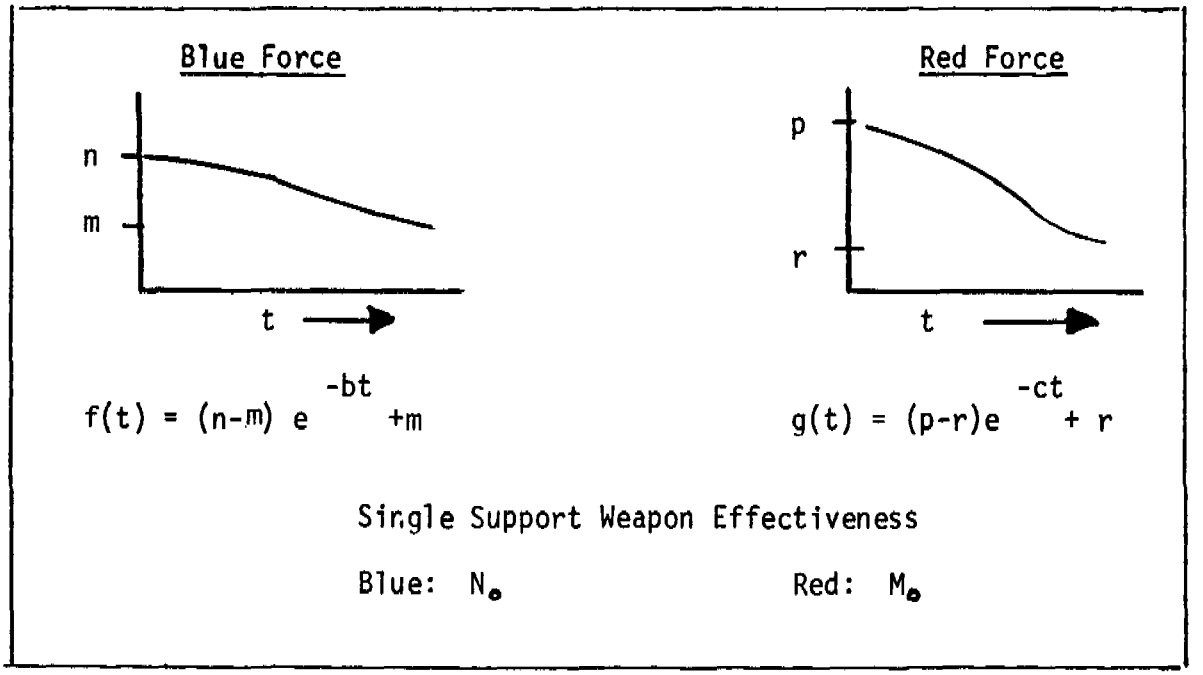


Figure 10. Support Weapon Employment Rate Model

$$\underline{\text{Blue}} \qquad \frac{dB}{dt} = -KR(t) - Mo\,(P - r)\,e^{-ct} - Mo\,r$$

$$\underline{\text{Red}} \qquad \frac{dR}{dt} = -JB(t) - No\,(n-m)\,e^{-bt} - No\,m$$

$$B(t) = \left(Bo + \frac{Nom}{J}\right)\cosh\left(t\sqrt{KJ}\right) - \left(Ro\sqrt{\frac{K}{J}} + \frac{Mor}{\sqrt{KJ}}\right)\sinh\left(t\sqrt{KJ}\right)$$

$$+ \left[\frac{No(n-m)\left(b\sqrt{\frac{K}{J}} - K\right)}{b^2 - KJ} + \frac{Mo(p-r)\left(\sqrt{KJ} - C\right)}{c^2 - KJ}\right]\sinh\left(t\sqrt{KJ}\right)$$

$$+ \frac{KNo\,(n-m)}{b^2 - KJ}\,e^{-bt} + \frac{CMo\,(p-r)}{c^2 - KJ}\,e^{-ct} - \frac{Nom}{J}$$

Figure 11. Loss Rates of Maneuver Elements

Table 6. Flexibility and Variations

1. Sensitivity analysis of uncertainties
2. Optimum tactics, employment, allocation
3. Quantity versus quality
4. Optimum mix: tradeoffs; break-even points
5. Adaptive tactics
6. Compare equal cost forces
7. Increasing effectiveness
8. Decreasing effectiveness
9. Increasing commitment
10. Decreasing commitment
11. Time-varying constraints
12. Diminishing returns
13. Factors within coefficients
14. Interdependence between coefficients
15. Non-linear relations
16. Alternative measures of effectiveness
17. Time period variations
18. Alternative levels of aggregation
19. Several echelons
20. Movement of forces
21. Lateral units
22. Model non-combat operations

HOMEOSTATIC CRITERIA FOR ASSESSMENT OF THE MORPHOLOGICAL STATE OF LARGE SCALE HYBRID ANALYTICAL-SIMULATION MODELS

Howard M. Bratt
Reliability and Maintainability
Modeling and Analysis Section, Eustis Directorate
US Army Air Mobility Research & Development Laboratory
Fort Eustis, Virginia 23604

A valid simulation data interval should represent an unbiased slice of time taken from a continuous process. Evaluating model achievement of stability is a process unique to each simulation model and to the simulation experiment under test. That is, a system has achieved a homeostatic condition when the major parameters by which the system performance is measured have reached a morphologically unchanging state.

This paper discusses the length of the stabilization period which should precede the data gathering portion of the simulation run (Fig 1). The simulation model under consideration is the Army's Aircraft Reliability and Maintainability Simulation model, a company level tactical operational model described in reference 1 and reference 2.

The analytical algorithms used in the stochastic ARMS Model to establish the approximate attainment of a morphological state (stability) are as follows:

1. Choose an acceptable stochastic parameter and sample its magnitude at even increments of time, e.g., once every day.

2. Put each sample in a table which also contains all prior samples up to this time.

3. As the ARMS model used General Purpose Simulation System (GPSS) Language, the current mean and standard deviation are readily available as each new sample is entered into the table.

4. Divide the table standard deviation by the square root of the current number of samples to obtain the standard error of the mean.

5. Carry on a simultaneous computation to obtain a "smooth exponential mean" utilizing a 20% tracking coefficient by summing 80% of the past value of the "smooth exponential mean" plus 20% of the latest sample. (This in turn, becomes the prior (80%) value for the next sample (20%)).

6. When the value of the smooth exponential mean falls within a band formed by the table mean $\pm$ the standard error obtained in (4) above, it is assumed that the model is stabilized.

This tracking algorithm is simple enough but its application requires considerable judgement and may vary between experimental scenarios. In the first place, it does not absolutely guarantee the existence of the morphological state for every stochastic parameter in the model. Only a set of algorithms monitoring every key stochastic variable could do that and the ARMS model's homeostatic criteria is monitoring only one stochastic parameter. Obviously, the choice of the parameter to be monitored is critical and it has taken considerable effort to arrive at the correct choice.

Some characteristics of the ARMS model require explanation in order that the reader understand the rationale for the choice of the stability criteria parameter. The ARMS model is close to being "universal" in the sense that it can accept almost any combination of number and type of aircraft, operational utilization flight frequencies, maintenance and repair policies, manpower, ground support equipment and logistics supply systems. The output statistics are readily convertible to operational cost estimates and mission performance effectiveness assessment.

Prior to start of a simulation run, each aircraft is initialized by assignment of a prior number of flight hours and calendar days to provide an approximately equal spacing between future flight hour a/o calendar day inspections (Fig. 2, 3). Highly critical components sometimes have a TBO (Time Between Overhaul) requirement and these components on each airplane are also initialized in relatively equal increments over their individual (each components TBO requirement may be different) TBO period. All of this initializing is very helpful in speeding up the simulation stabilization. However, when the simulation starts, no aircraft has any required maintenance actions and all are setting in the availability pool ready to go. This is not normal and until we have a "typical" number of maintenance actions distributed over the aircraft company we should not start collecting data.

Let us return once again to the discussion of some of the details of the ARMS model which will aid in the choice of the stability parameter. Airplane components fail in the ARMS model only during periods in which they are operating. Discovery of a component failure at the time it occurs is a probability assigned by the analyst to each element. If not discovered at times of failure, the probability of it being discovered at each subsequent event, e.g., flight, inspection, etc., is provided by the analyst. Up to this point we have identified THREE parameters of potential interest.

1. Failures which have occurred and were discovered at the time of occurrence.

2. Failures which have occurred and have not yet been discovered and,

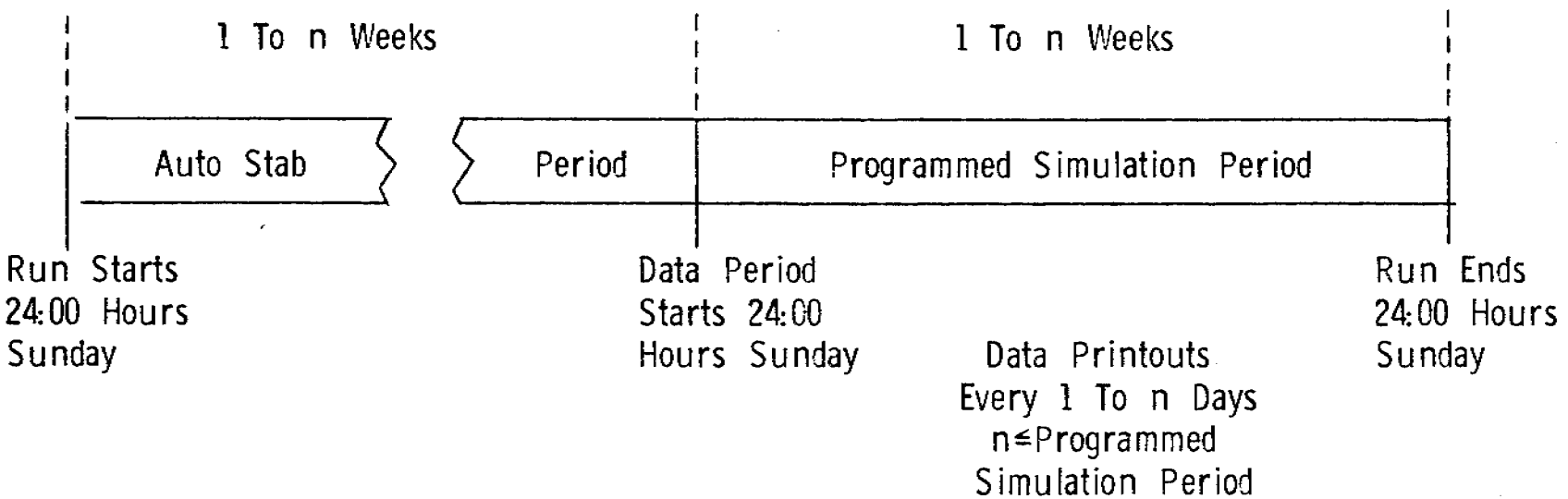
3. Failures which have occurred and were not discovered at time of occurrence but have subsequently been discovered.

Because of another feature of ARMS model policy, none of the above parameters can be used in their present form for our criteria. This feature is the categorizing of component failures into "Hard Down" (grounding the aircraft until repaired) and "Deferred Maintenance Actions" (fix as soon as possible but fly all required missions in the meantime). In addition, the ARMS analyst identifies the maximum number of "Deferred Maintenance Actions" he will permit on any one aircraft and any deferred maintenance actions in excess of this number will also send the aircraft "Hard Down." Because of this variability in the number of allowable deferred maintenance actions per aircraft in different simulation experiments, the daily total of deferred maintenance actions is not a good indication of stability. However, because both hard down maintenance actions and deferred maintenance actions eventually result in the assignment of maintenance actions to be worked off on the aircraft, this parameter "the number of maintenance actions assigned to be worked off in any 24 hour period" becomes the selected stability criteria parameter. By normalizing all stability computations to a constant number of aircraft (we chose 30) base, our stability criteria becomes independent of the number of aircraft in any experiment and our selected criteria for achievement of stability (although not the time required to achieve it) is independent of the aircraft utilization rates. We also conducted other investigations relative to the best value to use for the tracking coefficient prior to our selection of 80% of the last value plus 20% of the new sample.

When the stability criteria has been satisfied, the model automatically transitions from the stability period to the data gathering period on midnight of the following Sunday. At this time, many of the data values accumulated during the stability period are re-initialized (zeroed out) while other parameters remain unchanged (Fig. 4).

Again, there is no guarantee of stability with the selected criteria and the analyst must use reasonable judgement in the evaluation of the stability requirements of any given experiment. We have found that in a reasonably active simulation, this criteria will cause the model to run 3-4 weeks to attain stability. The length of the data run, after stabilization, is whatever was chosen by the systems analyst.

Figure 1



Automatic Stabilization Criteria

The Stabilization Period Lasts From 1 To n Weeks Until The Sampled Number of Deferred Maintenance Actions Is Determined Indigenously To Be From A Constant Distribution.

Programmed Stabilization Period

Lasts 1 To n Weeks As Exogenously Programmed

Figure 2

XB1 = 7 aircraft in model
X1 = 600 hr (36000 min) inspection cycle
X2 = 50 hr (3000 min) inspection sub-cycle
X3 = 10 hr (600 min) inspection sub-sub-cycle

$$PF1 = \frac{X1}{XB1} * [PB1 - 1] = \frac{36000}{7} \begin{bmatrix} 0 \\ \vdots \\ 6 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 30857 \end{bmatrix} \text{ minutes}$$

$$PF2 = X3 - [PF1 \text{ modulo } x\ 3] = 600 - \begin{bmatrix} 0 \\ \vdots \\ 257 \end{bmatrix} = \begin{bmatrix} 600 \\ \vdots \\ 343 \end{bmatrix}$$

$$PF3 = \frac{X3}{XB1} * PB1 = \frac{600}{7} \begin{bmatrix} 1 \\ \vdots \\ 7 \end{bmatrix} = \begin{bmatrix} 86 \\ \vdots \\ 600 \end{bmatrix}$$

$$PF1 + PF2 - PF3 = \begin{bmatrix} 8.6 \\ 87.1 \\ 175.7 \\ 254.3 \\ 342.8 \\ 421.4 \\ 510.0 \end{bmatrix} \text{ hours}$$

Figure 3

X1 Longest Flight Hour Inspection Cycle
X2 Interval Between Phases
X3 Interval Between Sub Phases

For Example:

A 600 Flight Hour Phased Inspection Cycle With The 12 Phases Occurring at 50 Hour Intervals and a Minimum Flight Inspection at 25 Hour Intervals Which Do Not Coincide With The Phased Inspection.

X1 = 600 Flight Hours
X2 = 50 Flight Hours
X3 = 25 Flight Hours

PHASED INSPECTION	0 1	0 2	0 3	0 4	0 5	0 6	0 7	0 8	0 9	0 10	0 11	0 12												
																								Repeat
	0	Δ	Δ	1	Δ	Δ	2	Δ	Δ	3	Δ	Δ	4	Δ	Δ	5	Δ	Δ	6	Flight
	0			0			0			0			0			0			0	Hours
	0			0			0			0			0			0			0	

0 1, 2, 3, ...12 - Phases of Phases Inspection
Δ - 25 Hour Inspection

Figure 4

REINITIALIZED		UNCHANGED	
EMT MEMT MMH	Sched-Unsched-NORS Combat-Reconfiguration Cannibalization		Cumulative Flight Hours Cumulative Calendar Days Day/Time Next Inspections Location--Pool-Maint-Etc. MA's Detected-Undetected Hard Down Maintenance
MISSION DATA	Called-Launched- Cancelled-Aborts Attrition-Consequences	AIRCRAFT STATUS	NORS-Cannibalized Parts Configuration Attrition Replacement Maintenance Actions In Work
AIRCRAFT DATA	Flights-Standby-Alert Reconfiguration-Ground Events Ready Pool-Availability		
MAINTENANCE MANPOWER BY MOS	Shift-Level-A--D Total-R/R-RIP-TBO Reconfig-Cannibalize Sched-Unsched-Combat Mission Events	LOGISTICS	Deterministic Inventory Status Parts On Order
LOGISTICS ELEMENT REPAIR DATA	Demand-Satisfaction-NORS Cannibalized-TBO Scrap-Repair-Repeat- At Levels A-B-C-D		
GSE	Called-Delays-Usage-Repaired		

References

1. Bratt, H. M., "Aircraft R&M Simulation (ARMS) Model," Proceedings 1975 Army Numerical Analysis Conference 11-12 February 1975, Eustis Directorate, US Army Air Mobility Research and Development Laboratory, Fort Eustis, Virginia.

2. The ARMS Report, Eustis Directorate, US Army Air Mobility Research and Development Laboratory, Fort Eustis, Virginia, TR-75-26A and TR-75-26B (to be published in July 75)

THE APPLICATION OF INFINITESIMAL TRANSFORMATION GROUPS TO THE SOLUTION OF NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS.

George W. Ullrich

US Army Mobility Equipment Research and Development Center
Fort Belvoir, Virginia 22060

ABSTRACT. The fundamental aspects of the method of infinitesimal transformation groups toward determining the global invariance group of partial differential equations are reviewed. The primary discussion rests upon the application of this method to a system of two nonlinear partial differential equations describing a biological diffusional transport process. With the full invariance group determined, we proceed to find invariant solutions for those equations under independent one-parameter subgroups. For this manifold of self-similar solutions the governing partial differential equations reduce to ordinary differential equations which involve only the similarity variables. The latter equations are solved explicitly for those cases which yield one-dimensional traveling wave solutions.

INTRODUCTION. In order to solve the nonlinear partial differential equations encountered in nature, the mathematician must choose prudently from among a variety of ad-hoc analytical methods of attack. While there is no general extant theory for nonlinear PDEs, some powerful and far-reaching methods of analysis have been advanced. In this discussion we present a group-theoretic method that may be employed to systematically generate self-similar solutions for both linear and nonlinear PDEs.

BACKGROUND. Consider the set of one-to-one continuous transformation which map the x-y space into itself,

$$x' = x'(x,y;\varepsilon) \qquad (1)$$
$$y' = y'(x,y;\varepsilon)$$

where ε denotes an arbitrary parameter. The transformations (1) comprise a group provided that:

1) Successive transformations satisfy a closure property

2) Each transformation has an inverse which is a member of the group

3) There exists an identity element $\hat{\varepsilon}$ so that

$$x'(x,y;\hat{\varepsilon}) = x \quad , \quad y'(x,y;\hat{\varepsilon}) = y \ .$$

Expanding the right-hand-side of Eqs. (1) about the identity transformation, denoted by $\varepsilon = \hat{\varepsilon}$, we obtain

$$\begin{aligned} x' &= x + (\varepsilon-\hat{\varepsilon})X(x,y) + O(\varepsilon-\hat{\varepsilon})^2 \\ y' &= y + (\varepsilon-\hat{\varepsilon})Y(x,y) + O(\varepsilon-\hat{\varepsilon})^2 \end{aligned} \tag{2}$$

Eqs. (2) constitute a group of infinitesimal transformations in the neighborhood for the identity transformation, i.e., they satisfy the aforementioned group properties to within $O(\varepsilon-\hat{\varepsilon})$.

Almost 100 years ago Lie (1881) began to develop his theory of one-parameter groups. And although he discussed the application of infinitesimal transformation groups to the solution of PDEs, subsequent work followed mostly along the path of finite groups. Most notable is the work of Dickson (1924) and Cohen (1931) who demonstrated the utility of group theory for integrating ordinary as well as partial differential equations. Recognizing the relationship between dimensional analysis and group theory, Birkhoff (1950) proposed a systematic analytical procedure for reducing the number of independent variables in a system of PDEs, thus facilitating integration. Morgan (1952) provided a generalization of this method with rigorous proof of its validity. In essence, the Morgan theory states that any system of PDEs, each of whose member equations is conformally invariant under a one-parameter group of transformation, may be reduced to a simpler system involving one less independent variable. This so called similarity reduction is effected by transforming the original n dependent and m independent variables to the set of n dependent and m-1 independent similarity variables which are the functionally independent group invariants. To each one-parameter group there corresponds a self-similar (invariant) solution which is obtained by solving the simpler reduced system.

For example, consider the steady-state heat conduction in a hollow right circular cylinder. The temperature distribution is governed by Laplace's equation

$$T_{xx} + T_{yy} = 0. \tag{3}$$

It is straightforward to show that Eq. (3) is conformally invariant (i.e., $T'_{x'x'} + T'_{y'y'} = \omega(x,y,\theta)[T_{xx} + T_{yy}]$) under the rotation group

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta \\ y' &= x\sin\theta + y\cos\theta \\ T' &= T \end{aligned} \tag{4}$$

where θ is the group parameter. Thus, the Morgan theory predicts a reduction in the PDE (3) to an ODE by introducing the similarity variables

$$\begin{aligned} R &= (x^2+y^2)^{1/2} \\ T &= T(R) \end{aligned} \tag{5}$$

which are the invariants of the group (4). Eq. (3) then becomes

$$\frac{d^2T}{dR^2} + \frac{1}{R}\frac{dT}{dR} = 0 \ . \tag{6}$$

One should note that the Morgan theory does not supply an effective method for determining all of the admissible one-parameter groups of transformations. A further deficiency is the lack of a formal mathematical procedure for establishing the group invariants. Consequently, in a typical application of the theory, simple one-parameter groups of transformation are assumed so that the invariants may be determined by inspection, as demonstrated above.

The latter shortcomings are eliminated by the use of infinitesimal groups as first shown by Ovsjannikov (1969). With this method all similarity variables and the functional forms of the concomitant self-similar solutions may be derived directly from the infinitesimal version of the full invariance group of transformations. Further applications and ramifications of the method are given by Bluman and Cole (1969), Woodard (1971), Ames (1972), Rosen and Ullrich (1973), and Ullrich (1974). In the following analysis we elucidate the fundamental aspects of the infinitesimal similarity method. However, the discussion will primarily focus on its application to a specific system of PDEs.

Consider the second-order system of nonlinear PDEs,

$$\frac{\partial\theta(x,t)}{\partial t} = x^{1-n}\,\frac{\partial\psi(x,t)}{\partial x} \tag{7}$$

$$\frac{\partial\psi}{\partial t} = x^{1-n}\theta^{\beta}\,\frac{\partial}{\partial t}(\theta^{-\beta}\,\frac{\partial\theta}{\partial x}) \ . \tag{8}$$

For the case of one-dimensional symmetry (n=1) and the identification,

$$\theta \equiv (\frac{s}{s_\infty})^{1-m} \quad , \quad 0 < m < 1 \tag{9}$$

$$\frac{\partial\theta}{\partial t} \equiv -(1-m)ks_\infty^{m-1}b \tag{10}$$

$$\beta \equiv \delta/(1-m)\mu \ . \tag{11}$$

Eqs. (7) and (8) become (Ullrich, 1974)

$$\frac{\partial b}{\partial t} = \mu\,\frac{\partial^2 b}{\partial x^2} - \delta\,\frac{\partial}{\partial x}(b\,\frac{\partial}{\partial x}\ell n s) \tag{12}$$

$$\frac{\partial s}{\partial t} = -ks^m b \ . \tag{13}$$

Letting b (x,t) and s (x,t) denote bacterial concentration and substrate agent concentration, respectively, and considering μ,δ and k as constant parameters, Eqs. (12) and (13) comprise a model for a biological diffusional transport process known as bacterial chemotaxis* (Keller and Segel, 1971). In particular, the latter system of equations describes the propagation of plane-symmetric bands of bacteria in a capillary containing a single substrate chemotactic agent that is consumed in the process. An extension of these equations to cases of two- and three-dimensional symmetry is given by Eqs. (7) and (8) with n=2 and n=3, respectively.

*Chemotaxis is the apparently purposeful response of some species of motile bacteria when subjected to gradients of a critical chemical in the substrate medium. The response is characterized by a spatial ordering (swarming) of bacteria into distinct bands (1-D) or rings (2-D) which move preferentially toward higher or lower concentrations of the substrate agent.

To recapitulate, the manifold of self-similar solutions to a system of PDEs is uniquely determined by the invariant solutions of those equations under appropriate one-parameter groups of transformations. Accordingly, we require Eqs. (7) and (8) to be simultaneously invariant under the generic one-parameter group of transformations

$$\begin{aligned} x' &= x'(x,t,u,v;\varepsilon) \\ t' &= t'(x,t,u,v;\varepsilon) \\ u' &= u'(x,t,u,v;\varepsilon) \\ v' &= v'(x,t,u,v;\varepsilon) \end{aligned} \tag{14}$$

where u and v are the generic dependent variables ($u = \theta(x.t)$ and $v = \psi(x,t)$ being unique solution surfaces), and ε is the group parameter.

In addition, we require the invariance of all boundary conditions $B_j(u,x,t) = 0$ & $C_j(v,x,t) = 0$ given on the curves $\beta_j(x,t) = 0$ and $\gamma_j(x,t) = 0$ that comprise the boundaries of the regions of definition of the unique solution surfaces $u = \theta(x,t)$ and $v = \psi(x,t)$, respectively. Note that the group of transformations (14) features complete mixing of the independent and dependent variables, thus defining a one-to-one mapping of the u, v, x, t space into itself.

According to the stipulated invariance conditions, the unique solution surfaces (associated with appropriately prescribed boundary conditions) must map into themselves, i.e., the solution surfaces must be invariant. Mathematically this is stated as

$$\begin{aligned} u'(x,t,\theta(x,t),\psi(x,t);\varepsilon) &= \theta(x',t') \\ v'(x,t,\theta(x,t),\psi(x,t);\varepsilon) &= \psi(x',t') \ . \end{aligned} \tag{15}$$

In principle, Eqs. (15) are sufficient to determine the admissible functional forms of θ and ψ.

We now make a decisive simplification, a la Lie, by studying the invariance in the neighborhood of the identity transformation. Consider the infinitesimal version of the group of transformations (14),

$$\begin{aligned} x' &= x + \varepsilon X(x,t,u,v) + O(\varepsilon^2) \\ t' &= t + \varepsilon T(x,t,u,v) + O(\varepsilon^2) \\ u' &= u + \varepsilon U(x,t,u,v) + O(\varepsilon^2) \\ v' &= v + \varepsilon V(x,t,u,v) + O(\varepsilon^2) \end{aligned} \tag{17}$$

where $\varepsilon=0$ corresponds to the identity transformation. It can be shown (see e.g. Cohen, 1931) that invariance of a function $f(x,y)$ under a

one-parameter, infinitesimal group of transformations is a necessary and sufficient condition for invariance of the function under the corresponding finite group. The advantage of using the infinitesimal approach in our analysis is that the invariance condition in the neighborhood of the identity implies linear (and hence tractable) equations for X, T, U, and V.

Using Eqs. (15) and (17) we proceed to write the infinitesimal versions of the invariance condition for the solution surfaces,

$$\begin{aligned} \theta(x+\varepsilon X, t+\varepsilon T) &= \theta(x,t) + \varepsilon U(x,t,u,v) + O(\varepsilon^2) \\ \psi(x+\varepsilon X, t+\varepsilon T) &= \psi(x,t) + \varepsilon V(x,t,u,v) + O(\varepsilon^2) \end{aligned} \tag{18}$$

Expanding the left hand side of Eq. (18) in a Maclaurin series in ε and equating terms of $O(\varepsilon)$ we find,

$$\begin{aligned} X(x,t,\theta,\psi)\theta_x + T(x,t,\theta,\psi)\theta_t &= U(x,t,\theta,\psi) \\ X(x,t,\theta,\psi)\psi_x + T(x,t,\theta,\psi)\psi_t &= V(x,t,\theta,\psi) \ , \end{aligned} \tag{19}$$

the equations describing the invariant surfaces.
The latter Eqs. (19) are first-order and quasi-linear with the associated Lagrange characteristic equations

$$\begin{aligned} \frac{dx}{X} &= \frac{dt}{T} = \frac{d\theta}{U} \\ \frac{dx}{X} &= \frac{dt}{T} = \frac{d\psi}{V} \ . \end{aligned} \tag{20}$$

We can find an integrating factor for the left equality of Eqs. (20) if X/T is independent of θ and ψ . The resulting integral

$$\eta(x,t) = \text{const.} \tag{21}$$

defines <u>similarity curves</u> in x, t space. From the right equalities of Eqs. (20) we then obtain

$$\begin{aligned} u(x,t) &= F_1(x,t,\eta,f_1(\eta)) \\ v(x.t) &= F_2(x,t,\eta,f_2(\eta)) \end{aligned} \tag{22}$$

where $f_1(\eta)$ and $f_2(\eta)$ are the solutions of the simultaneous ODEs obtained by substituting Eq. (22) into Eqs. (7) and (8). Hence, we have shown that the functional form of the similarity solution given by (22) can be derived directly from the infinitesimal version of the full invariance group.

We proceed to determine the admissible forms for the quantities X, T, U, V, i.e., those infinitesimal functions for which $\theta'(x',t')$ and $\psi'(x',t')$ are solutions to equations (7) and (8) with all the variables primed. This calculation requires expressions for the partial derivative terms appearing in equations (7) and (8) with the variables primed. Since $u = \theta(x,t)$ and $v = \psi(x,t)$ are the solution surfaces, it is expedient that all partial derivatives be calculated with respect to those surfaces the transformed space and time variables then being given by functions of the form

$$\begin{aligned} x' &= x'(x,t) \\ t' &= t'(x,t) \end{aligned} \quad . \tag{23}$$

From Eqs. (17) and (23) it follows that

$$\begin{aligned} \frac{\partial x}{\partial x'} &= 1 - \varepsilon[X_x + X_u\theta_x + X_v\psi_x] + O(\varepsilon^2) \\ \frac{\partial x}{\partial t'} &= - \varepsilon[X_t + X_u\theta_t + X_v\psi_t] + O(\varepsilon^2) \\ \frac{\partial t}{\partial t'} &= 1 - \varepsilon[T_t + T_u\theta_t + T_v\psi_t] + O(\varepsilon^2) \\ \frac{\partial t}{\partial x'} &= - \varepsilon[T_x + T_u\theta_x + T_v\psi_x] + O(\varepsilon^2) \quad . \end{aligned} \tag{24}$$

Using (24), we can now calculate the transformation equations that relate the various partial derivative terms. We first consider the transformation equations for u and v given by (17) along the solution surfaces θ and ψ and write

$$\begin{aligned} \theta'(x',t') &= \theta(x,t) + \varepsilon U(x,t,\theta,\psi) + O(\varepsilon^2) \\ \psi'(x',t') &= \psi(x,t) + \varepsilon V(x,t,\theta,\psi) + O(\varepsilon^2) \quad . \end{aligned} \tag{25}$$

Hence it follows that

$$\begin{aligned} \frac{\partial \theta'}{\partial t'} = \theta_t &+ \varepsilon\{U_t + (U_u - T_t)\theta_t - X_t\theta_x - T_u\theta_t^2 \\ &- X_u\theta_t\theta_x + U_v\psi_t - X_v\theta_x\psi_t - T_v\theta_t\psi_t\} + O(\varepsilon^2) \quad . \end{aligned} \tag{26}$$

and

$$\frac{\partial \psi'}{\partial t'} = \psi_t + \varepsilon\{V_t + (V_v - T_t)\psi_t - X_t\psi_x - T_v\psi_t^2 - X_v\psi_t\psi_x + V_u\theta_t - X_u\psi_x\theta_t - T_u\psi_t\theta_t\} + O(\varepsilon^2) \quad (27)$$

The spatial derivative terms $\partial\theta'/\partial x'$ and $\partial\psi'/\partial x'$ are formed from Eqs. (26) and (27) by interchanging the roles of t and x and T and X. Finally, the required mixed partial derivative term follows directly from (26)

$$\begin{aligned}\frac{\partial^2\theta'}{\partial x'\partial t'} = {} & \theta_{xt} + \varepsilon\{U_{xt} + (U_{xu} - T_{xt})\theta_t + (U_{ut} - X_{xt})\theta_x - T_{xu}\theta_t^2 \\ & - X_{ut}\theta_x^2 - T_{vu}\theta_t^2\psi_x + (U_{uu} - T_{ut} - X_{xu})\theta_x\theta_t - X_{uu}\theta_x^2\theta_t \\ & - X_{vu}\theta_t\psi_x\theta_x - T_{uu}\theta_t^2\theta_x - X_t\theta_{xx} - T_x\theta_{tt} + (U_u - X_x - T_t)\theta_{xt} \\ & - X_u\theta_{xx}\theta_t - T_u\theta_{tt}\theta_x - 2X_u\theta_{xt}\theta_x - 2T_u\theta_{xt}\theta_t + U_{xv}\psi_t \\ & + U_{vt}\psi_x - X_{vt}\psi_x\theta_x - T_{xv}\psi_t\theta_t + (U_{uv} - X_{xv})\psi_t\theta_x \\ & + (U_{vu} - T_{vt})\theta_t\psi_x - X_{vv}\psi_t\psi_x\theta_x - T_{vv}\psi_t\psi_x\theta_t + U_v\psi_{xt} \\ & - X_v\theta_{xx}\psi_t - T_v\theta_{tt}\psi_x - T_v\theta_{xt}\psi_t - T_v\theta_t\psi_{xt} - X_v\theta_x\psi_{xt} \\ & - X_v\psi_x\theta_{xt} - X_{uv}\theta_x^2\psi_t - T_{uv}\psi_t\theta_t\theta_x + U_{vv}\psi_t\psi_x\} + O(\varepsilon^2) .\end{aligned} \quad (28)$$

Rewriting Eqs. (7) and (8) in terms of the primed variables and substituting the latter transformation formulas for the derivative terms, we obtain

$$\frac{\partial\theta'}{\partial t'} - (x')^{1-n}\frac{\partial\psi'}{\partial x'} = \frac{\partial\theta}{\partial t} - x^{1-n}\frac{\partial\psi}{\partial x} + \varepsilon Q_1 + O(\varepsilon^2) \quad (29)$$

and

$$\begin{aligned}&\frac{\partial\psi'}{\partial t'} - (x')^{n-1}\left(\frac{\partial^2\theta'}{\partial x'\partial t'} - \beta(\theta')^{-1}\frac{\partial\theta'}{\partial x'}\frac{\partial\theta'}{\partial x'}\right) = \\ &\frac{\partial\psi}{\partial t} - x^{n-1}\left(\frac{\partial^2\theta}{\partial x\partial t} - \beta\theta^{-1}\frac{\partial\theta}{\partial x}\frac{\partial\theta}{\partial t}\right) + \varepsilon Q_2 + O(\varepsilon^2)\end{aligned} \quad (30)$$

where

$$Q_1 = U_t + (U_u - T_t)\theta_t - X_t\theta_x - T_u\theta_t^2 - X_u\theta_t\theta_x + U_v\psi_t - X_v\theta_x\psi_t$$
$$- T_v\psi_t\theta_t - (1-n)x^{-n}X\psi_x - x^{1-n}[U_x + (V_v - X_x)\psi_x - T_x\psi_t \qquad (31)$$
$$- X_v\psi_x^2 - T_v\psi_x\psi_t + V_u\theta_x - T_u\psi_t\theta_x - X_u\psi_x\theta_x]$$

and

$$Q_2 = V_t + (V_v - T_t)\psi_t - X_t\psi_x - T_v\psi_t^2 - X_v\psi_t\psi_x + V_u\theta_t - X_u\psi_x\theta_t$$
$$- T_u\psi_t\theta_t - (n-1)x^{n-2}X\theta_{xt} - x^{n-1}[U_{xt} + (U_{xu} - T_{xt})\theta_t + (U_{ut} - X_{xt})\theta_x$$
$$- T_{xu}\theta_t^2 - X_{ut}\theta_x^2 + (U_{uu} - T_{ut} - X_{xu})\theta_x\theta_t - X_{uu}\theta_x^2\theta_t - T_{uu}\theta_t^2\theta_x$$
$$- X_t\theta_{xx} - T_x\theta_{tt} + (U_u - X_x - T_t)\theta_{xt} - X_u\theta_{xx}\theta_t \qquad (32)$$
$$- T_u\theta_{tt}\theta_x - 2X_u\theta_{xt}\theta_x - 2T_u\theta_{xt}\theta_t + U_{xv}\psi_t + U_{vt}\psi_x - X_{vt}\psi_x\theta_x$$
$$- T_{xv}\psi_t\theta_t + (U_{uv} - X_{xv})\psi_t\theta_x + (U_{vu} - T_{vt})\theta_t\psi_x - X_{vv}\psi_t\psi_x\theta_x$$
$$- T_{vv}\psi_t\psi_x\theta_t + U_v\psi_{xt} - X_v\theta_{xx}\psi_t - T_v\theta_{tt}\psi_x - T_v\theta_{xt}\psi_t - T_v\theta_t\psi_{xt}$$
$$- X_u\theta_x\psi_{xt} - X_v\psi_x\theta_{xt} - X_{uv}\theta_x^2\psi_t - T_{uv}\psi_t\theta_t\theta_x + U_{vv}\psi_t\psi_x - T_{vu}\theta_t^2\psi_x$$
$$- X_{xu}\theta_t\psi_x\theta_x] + (n-1)x^{n-2}\beta X\theta^{-1}\theta_x\theta_t - x^{n-1}\beta\theta^{-2}U\theta_x\theta_t + x^{n-1}\beta\theta^{-1}\theta_t[U_x$$
$$+ (U_u - X_x)\theta_x + U_v\psi_x - T_x\theta_t - X_u\theta_x^2 - T_u\theta_t\theta_x - T_v\theta_t\psi_x - X_v\psi_x\theta_x]$$
$$+ x^{n-1}\beta\theta^{-1}\theta_x[U_t + (U_u - T_t)\theta_t - X_t\theta_x - T_u\theta_t^2 - X_u\theta_t\theta_x + U_v\psi_t$$
$$- X_v\theta_x\psi_t - T_v\psi_t\theta_t] .$$

Next, we ensure the invariance of the solution surfaces by requiring that the quantities Q_1 and Q_2 in Eqs. (29) and (30) be identically equal to zero. This so called "classical method" results in a system of linear partial differential equations that must be satisfied by the X, T, and V.

Since both θ and ψ can be prescribed arbitrarily at an initial instant of time, we must equate to zero the coefficients of the functionally independent terms in θ,ψ and their derivatives in $Q_1 = 0$ and $Q_2 = 0$. Now θ_t and ψ_t are specified by the right-hand side of Eqs. (7) and (8) and

must be eliminated by substition. the homogeneous coeffecient equations that ensue are linear in X, T, U, and V and can be satisfied by

$$\begin{aligned} X &= c_1 x + c_2 \quad ; \quad c_2(n-1) = 0 \\ T &= 2c_1 t + c_3 \\ U &= c_4 \theta \\ V &= [c_4 + (n-2)c_1]\psi . \end{aligned} \tag{33}$$

These latter equations comprise the infinitesimal version (Eqs. (17)) of the full invariance group of the system of equations (7) and (8). That is,

$$\begin{aligned} x' &= x + c_1 \varepsilon x + c_2 \varepsilon + O(\varepsilon^2) \quad ; \quad c_2(n-1) = 0 \\ t' &= t + 2c_1 \varepsilon t + c_2 \varepsilon + O(\varepsilon^2) \\ \theta' &= \theta + c_4 \varepsilon \theta + O(\varepsilon^2) \\ \psi' &= \psi + [c_4 + (n-1)c_1]\varepsilon\psi + O(\varepsilon^2) \end{aligned} \tag{34}$$

where c_1, c_2, c_3, and c_4 are arbitrary constants.

The corresponding finite group of transformations (which for this simple case may be written by inspection) may be derived formally by using the Lie group generators. The result is stated with a convenient redefinition of the translation parameters as

$$\begin{aligned} x' &= e^{\alpha} x + \kappa \; ; \; \kappa(n-1) = 0 \; ; \; \kappa \equiv \frac{c_2}{c_1}(e^{c_1 \varepsilon} - 1) \\ t' &= e^{2\alpha} t + \lambda \; ; \; \alpha \equiv c_1 \varepsilon \; , \; \lambda \equiv \frac{c_3}{2c_1}(e^{2c_1 \varepsilon} - 1) \\ \theta' &= e^{\gamma} \theta \; ; \; \gamma \equiv c_4 \varepsilon \\ \psi' &= e^{[\gamma + (n-2)\alpha]} \psi \; . \end{aligned} \tag{35}$$

Due to the arbitrariness of the constants c_1, c_2, c_3, and c_4, the latter transformations comprise a 4-parameter group, displaying translatinal invariance in t, translational invariance in x (restricted to the one-dimensional version of the governing equations), similitudinous scale invariance in t and x, and scale invariance in the dependent variables θ and ψ. Clearly, the transformations (34) have ∞^4 one-parameter subgroups. Using the real 4-tuple $\underset{\sim}{c} = (c_1, c_2, c_3, c_4)$ as a label, each one-parameter subgroup is designated by a fixed $\underset{\sim}{c}$ with the common subgroup parameter ε ranging over the entire real number scale.

We proceed to determine the classes of one-parameter subgroups that have qualitatively different group-space trajectories and consequently lead to functionally independent self-similar solutions to the original system nonlinear PDEs.

It is convenient at this point to combine the original second-order system of equations (Eqs. (7) and (8)) yielding the single third-order nonlinear PDE

$$\frac{\partial^2\theta}{\partial t^2} = \mu x^{1-n} \frac{\partial}{\partial x}[x^{n-1}\theta^{\beta} \frac{\partial}{\partial t}(\theta^{-\beta} \frac{\partial\theta}{\partial x}] \tag{36}$$

with the associated invariance group of infinitesimal transformations

$$\begin{aligned} x' &= x + c_1\varepsilon x + c_2\varepsilon + O(\varepsilon^2) \\ t' &= t + 2c_1\varepsilon t + c_2\varepsilon + O(\varepsilon^2) \\ \theta' &= \theta + c_4\varepsilon\theta + O(\varepsilon^2) \end{aligned} \tag{37}$$

For the class of one-parameter subgroups designated by $c_1 \neq 0$, the Lagrange characteristic equation (20) becomes

$$\frac{dx}{c_1 x + c_2} = \frac{dt}{2c_1 t + c_3} = \frac{d\theta}{c_4\theta}$$

Integrating, we obtain the subgroup invariants

$$\eta = \frac{(x + c_2/c_1)^2}{t + c_3/2c_1}$$

$$f = \frac{\theta}{(t + c_3/c_1)^{c_4/2c_1}} \tag{38}$$

which are, in fact, the integration constants. Thus, by Eq. (22), the corresponding general subgroup invariant equation for θ, i.e., the self-similar solution, takes the form $f = f(\eta)$ or equivalently,

$$\theta = (t + c_3/c_1)^{c_4/c_1} f(\eta) \tag{39}$$

The second class of one-parameter subgroups are designated by $c_1 = 0$, and lead to the Lagrange characteristic equation

$$\frac{dx}{c_2} = \frac{dt}{c_3} = \frac{d\theta}{c_4\theta} \tag{40}$$

With the further restriction of $c_2 \neq 0$, $c_3 \neq 0$, we integrate the left and right equations in (40) and again obtain two integration constants which are the functionally independent subgroup invariants

$$\eta = x - (c_2/c_3)t$$
$$f = \exp[-(c_4/c_3)t]\theta \ . \tag{41}$$

The functional form of the corresponding self-similar solution for θ is then given by

$$\theta = e^{(c_4/c_3)t} f(\eta) \ . \tag{42}$$

Eqs. (39) and (42) comprise the manifold of self-similar solutions to Eq. (36). By substituting the latter expressions into Eq. (36) we obtain an ordinary differential equation for $f(\eta)$ in each case. In the following paragraphs we restrict our attention to the wave solutions (42) which are of more immediate physical interest in view of the existing experimental data. Specific solutions of the type (39) are given elsewhere (Ullrich, 1974).

Defining $c_2/c_3 = c$ as the constant propagation speed, Eq. (42) describes steadily propagating waves that are uniform $(c_4/c_3 = 0)$ or exhibit exponential decay $(c_4/c_3 < 0)$ or growth $(c_4/c_3 > 0)$. Since we require $c_2 \neq 0$, the condition $(n-1)c_2 = 0$ in (34) restricts the latter wave solutions to the special one-dimensional, plane-symmetric case of the governing dynamical equation (36). Furthermore, it is noteworthy that accelerative wave solutions which exhibit a time-dependent propagation velocity are precluded by Eq. (42). This is directly attributable to the fact that the invariance group (34) does not admit independent scale transformations in x and t.

We will consider two cases of Eq. (42) denoted by $c_4 = 0$ and $c_4 \neq 0$. For the former case, Eq. (42) reduces to

$$\theta = f(\eta) \ , \ \eta = x - ct \ ,$$

which when substituted into Eq. (36) yield the ordinary differential equation

$$\mu c f''' + c^2 f'' + \frac{\mu\beta c f'^2}{f^2} - \frac{2\mu\beta c f'' f'}{f} = 0 \ . \tag{43}$$

Supplemented with the boundary conditions*

$$f \to 0 \ , \quad \eta \to -\infty$$

$$f \to 0 \ , \quad \eta \to +\infty$$

$$f' \to 0 \ , \quad \eta \to +\infty$$

appropriate for the Adler band experiments (Adler, 1969), the solution to equation (43) is given by

$$\theta = f(\eta) = (1 + e^{-c\eta/\mu})^{-1/(\beta-1)} \ , \quad \beta > 1 \tag{44}$$

With the definitions (9) through (11), equation (44) begets the explicit expressions for the bacterial density and substrate concentration distributions

$$b = \frac{c^2 s_\infty^{1-m}}{k[\delta - (1-m)\mu]} e^{-c\eta/\mu} (1 + e^{-c\eta/\mu})^{-\delta/[\delta - (1-m)\mu]} \tag{45}$$

$$s = s_\infty (1 + e^{-c\eta/\mu})^{-\mu/[\delta - (1-m)\mu]} \tag{46}$$

where $$[1 - \delta/\mu] < m < 1. \tag{47}$$

Eqs. (45) and (46) have been shown to be in good agreement with experimental data (Keller and Segel, 1971; Rosen, 1974).

For the case $c_4 \neq 0$, we consider the general form of the self-similar solution (42),

$$\theta = e^{pt} f(\eta) \ , \quad \eta = x - ct$$

---

*Similarity reductions generated by the Morgan theory become physically useful only if the similarity solutions are compatible with the original auxiliary conditions. This compatibility requirement can be tested by checking a posteriori whether the auxiliary conditions are expressible, without inconsistency, in terms of the similarity variables. This will not always be the case. For example, the similarity reduction of a PDE in three variables with two boundary conditions and one initial condition requires that the original auxiliary conditions coalesce to form two boundary conditions on the independent similarity variable.

where $p \equiv c_4/c_3$. Unfortunately, this functional form prevents satisfaction of a boundary condition germane to the one-dimensional, plane-symmetric, bacterial-band propagation experiments (see Adler, 1969), viz.,

$$\theta \to 1 \quad , \quad \eta \to \infty \quad ; \quad t \geq 0$$

Nonetheless, traveling wave solutions that exhibit exponential growth are intuitively physical for the model in question. In fact the classic example of exponential biological growth is bacterial growth in an unlimited culture medium where overcrowding effects are precluded. Thus we propose the alternate boundary conditions

$$\begin{aligned} f(\eta) &\to 0 \quad , \quad \eta \to -\infty \\ f(\eta) &\to 0 \quad , \quad \eta \to +\infty \end{aligned} \quad ; \quad t \geq 0$$

which are suggestive of an experiment wherein a bacterial band originates at each end of a long capillary so that both bands propagate toward each other, depleting all of the substrate chemotactic agent behind them. (Such an experiment has not been performed to date). The ODE for $f(\eta)$ that results is given by

$$\mu c f''' + [c^2 + \mu(\beta-1)p]f'' - 2cpf' + p^2 f = \frac{2\mu\beta c f'' f'}{f} - \frac{\mu\beta c f'^3}{f^2} \tag{48}$$

Eq. (48) may be reduced to the first-order ODE

$$u' - \frac{u}{z} - \{(3-2\beta) + \frac{1}{z}[\frac{c^2 + (\beta-1)p}{\mu c}]\}u^2 - \{(1-\beta)z + [\frac{c^2 + (\beta-1)p}{\mu c}] - \frac{2p}{\mu z} + \frac{p^2}{\mu c z^2}\}u^3 = 0 \tag{49}$$

by successively imposing the variable transformations

$$\begin{aligned} y(f) &\equiv f'(\eta) \\ z(f) &\equiv y/f \\ q(z) &\equiv fz' \\ u(z) &\equiv q^{-1} \end{aligned}$$

Eq. (49) is an Abel equation of the first kind which for small values of p (large β) has as its formal solution

$$\int_{u_o}^{u} u\left[\frac{\gamma(a+1)(3-2\beta)}{2a} z^2 + \frac{2ap\gamma}{\mu(3-2\beta)} + \gamma u\right]^a du$$
$$+ \int_{z_o}^{z} \left[(1-\beta)z^3 - \frac{2pz}{\mu} + (3-2\beta z u_o\right] X$$
$$\left[\frac{\gamma(a+1)(3-2\beta)z^2}{2a} + \frac{2ap\gamma}{\mu(3-2\beta)} + \gamma u_o\right]^a dz = \text{const.} \qquad (50)$$

where γ is an arbitrary constant, a is a constant whose admissible values are prescribed by

$$a_{\pm} = \frac{-\ell \pm \sqrt{\ell^2 - 4\ell}}{2} \quad , \quad \ell \equiv \frac{(3-2\beta)^2}{2(1-\beta)}$$

and u_o and z_o are constants that may be chosen as is convenient.

CONCLUSION. In conclusion, we have demonstrated the utility of infinitesimal transformation groups toward systematically arriving at the manifold of self-similar solutions for PDEs. The method is particularly suitable for nonlinear PDEs because the group-theoretic approach entails no implicit assumptions of linearity. On the other hand, the richness of the manifold of self-similar solutions is dependent on the degree of symmetry exhibited by the governing system of equations. When higher order nonlinearities are present, the probability of uncovering subtle symmetries decreases substantially. In the final analysis, the usefulness of the approach advanced in this paper must be weighed intuitively for each application.

REFERENCES

Adler, J. (1969). Science, N.Y. 166, 1588.

Ames, W.F. (1972). Nonlinear Partial Differential Eqs. in Engineering, Vol. II, Academic Press, N.Y.

Birkhoff, G. (1950). Hydrodynamics, 1st ed., Princeton University Press, Princeton, N.J.

Bluman, G.W. & Cole, J.D. (1969). J.Math. Mech. 18, No. 11,1025.

Cohen, A. (1931). An Introduction to the Lie Theory of One-Parameter Groups, Stechert, N.Y.

Dickson, (1924). Ann. Math. (2) 25, 287.

Keller, E.S. & Segel, L.A. (1971). J. Theor. Biol. 30. 225.

Lie, S. (1881). Math. (Kristiana) 6, 328.

Morgan, A.J.A. (1952). Quart. J. Math., Oxford Ser. 2, 250.

Ovsjannikov, L.V. (1969). Soviet Math. Dokl. 10, 538.

Rosen, G. & Ullrich, G.W. (1973). SIAM J. Appl. Math. 24, #3, 286.

Ullrich, G.W. (1974). Ph.D. Dissertation, Drexel U., Phila., Pa.

Woodard, H. , Jr. (1971). Ph.D. Dissertation, University of Iowa.

ON UNIQUENESS OF PIECEWISE POLYNOMIAL APPROXIMATION

C. K. CHUI, P. W. SMITH, and J. D. WARD
Department of Mathematics
Texas A&M University

ABSTRACT. This paper deals with uniform approximation of continuous functions by piecewise polynomials. A uniqueness theorem for placing the knots which balance the error is established. Counter-examples are also given to show nonunicity for many functions. Numerical examples are given to demonstrate the efficiency of piecewise polynomial approximation for certain functions.

1. INTRODUCTION. Although the ideas involved in piecewise polynomial approximation are quite elementary, it has been only recently that much effort has been expended in studying these approximants. In particular, Burchard [1], Rice [5], and Handscomb [2] among others have studied various facets of the field.

If we denote by P_k^n the set of functions which have k polynomial pieces of order n (degree $< n$), then one of our results is: Let $f \in C[a,b]$ satisfy $f^{(n-1)}$ is strictly monotone. Then f has a unique best uniform approximant from $P_k^n[a,b]$. A weaker version of this result is stated in [3]. We will give an independent and complete proof of our result. We also remark that Handscomb [2] has the following result which is closely related to this above theorem.

Theorem (Handscomb). If $f^{(n-2)}$ exists and is strictly convex in $[a,b]$, then the best uniform approximation to f on $[a,b]$ by a spline of degree $(n-1)$ $(n \geq 3)$ with k free knots is unique, all of its knots are genuine and distinct, and all its coefficients β_j are positive.

Notice that Handscomb's condition ($f^{(n-2)}$ strictly convex) is essentially the same as $f^{(n-1)}$ strictly monotone. Furthermore, $P_k^n \subset S_{kn}^n$ where S_{kn}^n is the set of splines of order n with kn knots. It follows that $\operatorname{dist}(f,S_{kn}^n) \leq \operatorname{dist}(f,P_k^n)$. However, we believe that it is much easier to compute the best approximant to f from P_k^n than from S_{kn}^n. This computational advantage we feel justifies further research into piecewise polynomial approximation.

2. UNIQUE BALANCED PARTITIONS. A mesh of order k is a non-decreasing $(k+1)$-tuple of real numbers or alternatively it is a collection of k open intervals $I_i = (\mu_{i-1},\mu_i)$, $i = 1, \ldots, k$. If $u = (u_0, \ldots, u_k)$ is a mesh, let k denote the order and $\lambda(u) =$

$\max_{1\le i\le k} (u_i - u_{i-1})$ the mesh-size of u. A partition is a mesh with nonempty intervals. If u is a mesh on the open interval (a,b), i.e., $u_0 = a$ and $u_k = b$, then $P^n(u)$ is a collection of real functions on (a,b) whose restrictions to the intervals of u are polynomials of degree at most $n-1$. P_k^n will denote the set of all piecewise polynomials of degree at most $n-1$, whose knots form a kth order mesh. If $f \in L^p(a,b)$, $1 \le p \le \infty$, let

$$E_{p,n}(f,u) = \inf \{\| f - s\|_{p,(a,b)} : s \in P^n(u)\}$$

where u is a mesh on (a,b) and let

$$E_{p,n}(f,k)_{(a,b)} = \inf_u \{E_{p,n}(f,u): u_0 = a, u_k = b\} .$$

By a balanced mesh u of order k, we mean a mesh u having the property that $E_{p,n}(f,I_j) = E_{p,n}(f,I_i)$ for all I_i and $I_j \in u$. By an optimal mesh u of order k, we mean a mesh u which satisfies $E_{p,n}(f,k) = E_{p,n}(f,u)$.

In [1], Burchard studied piecewise polynomial approximation on optimal meshes. His primary interest was to determine bounds and asymptotic limits for the error function $E_{p,n}(f,k)$.

The purpose of this paper is to exhibit a fairly general class of functions having a unique best approximant from P_k^n in $L^\infty(a,b)$. We first derive an important relationship between balanced error and optimal partitions which has special computational implications.

Proposition 1. In $L^\infty(a,b)$, the balanced error partition exists and is an optimal partition.

Proof. The existence of the balanced error partition is a consequence of Lemma 3.2 of [1]. Now suppose the balanced partition $u = (t_0, \ldots, t_k)$ is not optimal. Let $u_1 = (\tau_0, \ldots, \tau_k)$ be another partition for which $E_{\infty,n}(f,u_1) < E_{\infty,n}(f,u)$. Since u is balanced, it follows that $\tau_1 < t_1$. Similarly, $\tau_2 < t_2$, $\ldots$, $\tau_{k-1} < t_{k-1}$. But clearly, this is impossible since $E_{\infty,n}(f,(\tau_{k-1},b)) \ge E_{\infty,n}(f,(t_{k-1},b)) = E_{\infty,n}(f,u)$. This completes the proof.

We now give an example which shows that in spaces other than $L^\infty(a,b)$ a balanced partition need not be optimal.

Example. Our approximating manifold in $L^1[0,4]$ is the piecewise constant functions with one variable knot. Let f be defined as

$$f(x) = \begin{cases} 1 & x \in [0,2] \\ 2 & x \in [2,3] \\ h(x) & x \in [3,4] \end{cases}$$

where $h(x)$ has the properties that $C = 2$ is a best constant approximation to $h(x)$ on $[3,4]$ and $\| h(x) - 2 \|_1 = 4/3$. Since $C = 4/3$ is a best constant approximant to $f(x)$ on $[0,3]$, the knot at $x = 3$ gives rise to a balanced error partition $u = (0,3,4)$ and $E_{1,1}(f,u) = 8/3$. On the other hand, by choosing the knot at $x = 2$, we can select our best approximant to be $y = 1$ on $[0,2]$ and $y = 2$ on $[2,4]$ with total error $4/3$. Thus, the balanced error partition is not optimal. We note that we could have made $f(\cdot)$ continuous and still obtained the same result with only slight modifications in the above example. We are now ready to state

Theorem 1. If $f \in C[a,b]$ and $f^{(n-1)}$ is strictly monotone on (a,b), then f has a unique best approximant from P_k^n.

The proof of this theorem follows directly from the following three lemmas.

Lemma 1.1. If f is as in Theorem 1, $a < t_1 < t_2 < b$ and p^* is the (unique) best nth order approximating polynomial to f on $[t_1,t_2]$ then

$$|p^*(t_i) - f(t_i)| = \| p^* - f \|_{[t_1,t_2]} , \quad i = 1, 2 .$$

Proof. Suppose to the contrary, that

$$|p^*(t_1) - f(t_1)| < \| p^* - f \|_{[t_1,t_2]} . \tag{1.1}$$

Then by the Chebyshev Alternation Theorem, there exist points $\tau_1, \cdots, \tau_{n+1}$ with $t_1 < \tau_1$ at which

$$p^*(\tau_i) - f(\tau_i) = (-1)^i [\pm \| p^* - f \|_{[t_1,t_2]}]$$

and $\tau_i < \tau_{i+1}$, $i = 1, \ldots, n$. Further, by (1.1) we see that $(p^* - f)'(\tau_i) = 0$, $i = 1, \ldots, n$. Thus, by Rolle's theorem applied $n - 2$ times we see that there are two points ε_1 and ε_2

in (t_1,t_2) so that

$$(p^* - f)^{(n-1)}(\varepsilon_i) = 0 \quad , \quad i = 1, 2 \quad .$$

But this is clearly a contradiction since $(p^*)^{(n-1)}$ is a constant and $f^{(n-1)}$ is assumed to be strictly monotone. Of course, the argument is symmetric if in (1.1) we replace t_1 by t_2 and this completes the proof of Lemma 1.

We now obtain two corollaries of Lemma 1 which are needed later.

Corollary 1.1. The functional $E_{\infty,n}(f,(t_1,t_2))$ is a strictly increasing function in t_2 and a strictly decreasing function in t_1.

Proof. Recall that in the proof of Lemma 1, $(f - p^*)'$ cannot vanish at either t_1 or t_2. Thus, the assertion of Corollary 1 follows.

Corollary 1.2. $f - p^*$ attains its norm exactly $n + 1$ times in $[t_1,t_2]$.

Lemma 1.2. If f is as in Theorem 1, then $E_{\infty,n}(f,k) > E_{\infty,n}(f,k+1)$.

Proof. Let $t_0 < \ldots < t_k$ be a balanced (hence optimal) error partition. Then by adding a new break point, say τ, between a and t_1, we note that

$$E_{\infty,n}(f,(a,\tau)) < E_{\infty,n}(f,(a,t_2)) = E_{\infty,n}(f,k) > E_{\infty,n}(f,(\tau,t_1)) \ .$$

by Corollary 1. Hence $E_{\infty,n}(f,(\tau,t_1+\varepsilon_1)) < E_{\infty,n}(f,(a,t_2))$ for small enough positive ε_1. But then, appealing to Corollary 1 again,

$$E_{\infty,n}(f,(t_1+\varepsilon_1,t_2)) < E_{\infty,n}(f,(a,t_1))$$

and so there exists an $\varepsilon_2 > 0$ such that

$$E_{\infty,n}(f_1,(t_1+\varepsilon_1,t_2+\varepsilon_2)) < E_{\infty,n}(f,(a,t_1)) \quad .$$

Continuing in this manner, we produce a new partition $\{t_0,\tau,t_1+\varepsilon_1, \ldots, t_{n-1}+\varepsilon_{n-1},t_n\} = \pi$ so that

$$E_{\infty,n}(f,k+1) \leq E_{\infty,n}(f,\pi) < E_{\infty,n}(f,k) \ .$$

This completes the proof.

Lemma 1.3. The only optimal partition is the balanced error partition, if f is as in Theorem 1.

Proof. Let $u = (t_0, \ldots, t_k)$ be an optimal balanced partition for f. Suppose that there is another partition $u_1 = (\tau_0, \ldots, \tau_k)$ which is also optimal for f. Then by Corollary 1, it follows that $\tau_1 \leq t_1, \ldots, \tau_{k-1} \leq t_{k-1}$. Reapplying Corollary 1, one obtains that $\tau_{k-1} \geq t_{k-1}, \ldots, \tau_1 \geq t_1$. This proves the lemma.

Proof of Theorem 1. Lemma 3 guarantees the existence of a unique optimal mesh, namely the balanced error partition $u = (\tau_0, \ldots, \tau_k)$. Since on each interval $[\tau_i, \tau_{i+1}]$ $i = 0, \ldots, k-1$ the best approximating polynomial of order n is unique, the best piecewise polynomial approximation p^* is unique. We require p^* to be right continuous at each τ_i to make p^* well defined at the knots. The proof is completed.

The following is a consequence of Theorem 1.

Corollary 1.3. If f satisfies the hypotheses of Theorem 1 and $k \geq 1$, then the (unique) best approximation to f from P_k^n is continuous if and only if n is even.

Proof. By Lemma 1, the error $\| f - p^* \|$ is attained at the knots. Thus, one only needs to check whether $p^*|_{[t_{i-1},t_i]}$ and $p^*|_{[t_i,t_{i+1}]}$ match up at t_i or are on opposite sides of f.

3. NUMERICAL PROCEDURES. In order to find the optimal knots of the best piecewise polynomial approximant to f, we tried several variants of schemes suggested in [4]. The method suggested there is a fixed point iteration scheme. More specifically let $u^{(m)} \equiv (u_0^{(m)}, \ldots, u_k^{(m)})$ be the partition at the mth step in the iteration procedure. One then may adjust the knots according to many criteria. We compare three methods.

$$\text{(I)} \quad u_i^{(m+1)} = u_i^{(m)} + C[E_{\infty,n}(u_i, u_{i+1}) - E_{\infty,n}(u_{i-1}, u_i)]$$

$$i = 1, \ldots, k-1.$$

$$\text{(II)} \quad \text{Set:} \quad A_i^{(m)} = (1/i) \sum_{j=1}^{i} E_{\infty,n}(u_{j-1}^{(m)}, u_j^{(m)})$$

$$\text{Then set:} \quad u_i^{(m+1)} = u_i^{(m)} + c_i i (A_k^{(m)} - A_i^{(m)})$$

$$i = 1, \ldots, k-1$$

$$\text{(III)} \quad \text{Set:} \quad B_i^{(m)} = (1/i) \sum_{j=n-i+1}^{n} E_{\infty,n}(u_{j-1}^{(m)}, u_j^{(m)})$$

$$\text{Set: } V_i^{(m)} = u_i^{(m)} + c_i i(-B_k^m + B_i^m)$$

$$Z_i^{(m)} = u_i^m + c_i' i(A_k^{(m)} - A_i^{(m)})$$

$$i = 1, \ldots, k-1$$

$$\text{Set: } u_i^{(m+1)} = (V_i^{(m)} + z_i^{(m)})/2 \; ; \; i = 1, \ldots, k-1.$$

In all the above schemes, it is assumed that the numbers c_i $i = 1, \ldots, k-1$ are positive. It is clear that the balanced error solution is a fixed point of all the above schemes. In [4] it is shown that these schemes will converge to a fixed point provided the c_i are chosen small enough.

Specifically, if $c_i = c < (1/M)[u_k - u_0]$ where M is the Lipshitz constant of $E_{\infty,n}(t_1, \cdot)$, then the methods I, II, and III converge. However, the choice of the c_i is delicate since too small a c_i will converge very slowly and too large a c_i will fail to converge.

We found that a convenient conservative estimate for c_i, or c_i' is $|(p_i^* - f)'(u_i)|^{-1}$ where p_i^* is the best polynomial approximant to f on (u_i, u_{i+1}).

A simple Remes exchange algorithm is used to compute an approximation to the best polynomial approximant to f on each of the subintervals. Our experience has been that method III generally produces the best results. This should not be surprising since it adjusts the knots according to the global error (as does II) and secondly, it is not dependent on whether one starts from the right or left endpoint as is method II.

The following table typically illustrates the behavior of the methods. This table represents the results in trying to approximate $F(x) = |x|^3$ on $[-3,2]$ by piecewise linear polynomials.

Method	Number of Knots	Iterations	Error
I	1	37	2.556
II	1	201	2.561
III	1	91	2.556
I	2	106	1.2929
II	2	201	1.2935
III	2	117	1.2929
I	3	201	0.6367
II	3	247	0.6376
III	3	83	0.6377
I	4	420	0.4287
II	4	337	0.4274
III	4	102	0.4303

These results are quite representative of the relationships between the methods. In general, for 1 knot (independent of degree) method I appears to be much faster than II or III. Method III nearly always beats method II and after 1 or 2 knots easily outdistances the other methods.

The authors would like to express their gratitude to Robert Strader who did much of the programming.

REFERENCES

1. H. Burchard, Splines (with Optimal Knots) Are Better, J. Applicable Analysis, to appear.

2. D. C. Handscomb, Characterization of Best Spline Approximations with Free Knots, Approximation Theory, A. Talbot (ed.), Academic Press, New York, p. 63 - p. 70.

3. G. Meinardus, Approximation of Functions: Theory and Numerical Methods. Springer Tracts in Natural Philosophy, Vol. 13, Springer-Verlag New York, 1967, p. 188.

4. T. Pavlidis and A. P. Maika, Uniform Piecewise Polynomial Approximation with Variable Joints, J. of Approximation Theory, Vol. 12, (1974) pp. 61-69.

5. J. R. Rice, On the Degree of Convergence of Nonlinear Spline Approximation, Approximation with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Acad. Press, New York, 1969, p. 349 - p. 365.

Linear Generalizations of the Gronwall-Reid-Bellman Lemma

Jagdish Chandra Paul Wm. Davis

U. S. Army Research Office
Research Triangle Park, North Carolina

An investigation of uniqueness or continuous dependence on boundary data for the system of nonlinear equations,

$$(1) \qquad w_{xy} = F(x, y, w),$$
$$w \text{ prescribed on } x = x^{\circ}, \; y = y^{\circ},$$

where w is a vector and F is a matrix which is Lipschitz continuous in w, reduces to a study of integral inequalities of the general form

$$(2) \qquad u(x, y) \leq a(x, y) + G(x, y) \int_{x^{\circ}}^{x} \int_{y^{\circ}}^{y} H(s, t)u(s, t)\, ds\, dt.$$

Here a is a vector determined by the boundary conditions and G and H are continuous matrices with nonnegative entries.

Inequality (2) is an analog for systems in several independent variables of the Gronwall-Reid-Bellman inequality used in the theory of ordinary differential equations, e.g. [P. Hartman, Ordinary Differential Equations, Wiley, p. 24]. By manipulating the resolvent

kernel of the integral operator on the right side of (2), we have obtained the upper bound required to establish uniqueness, continuous dependence, etc. for the system (1) of partial differential equations:

Theorem: If $u(x, y)$ satisfies (2), then

$$u(x, y) \leq a(x, y) + G(x, y) \int_{x^0}^{x} \int_{y^0}^{y} V(x, y; s, t)H(s, t)a(s, t)dsdt,$$

where V is defined by

$$V(x, y; s, t) = I + \int_{s}^{x} \int_{t}^{y} H(p, q)G(p, q)V(p, q; s, t)\, dp\, dq.$$

The restriction to merely two variables is not essential.

The proof of this theorem is simpler and the result more general than that of D. R. Snow [Proc. Amer. Math. Soc. 33 (1972), 46-54] and E. C. Young [ibid. 41 (1973), 241-244], who used a differential-initial-value approach. In several special cases, the matrix V can be computed explicitly, thereby recovering a number of earlier results, including the classical, single-variable, scalar lemma of Gronwall, Reid, and Bellman.

The discrete analog of (2), which arises in the numerical solution of the system (1) of partial differential equations by Euler's method, can be treated similarly.

A METHOD FOR THE NUMERICAL SOLUTION OF TWO-POINT BOUNDARY VALUE PROBLEMS BASED ON THE USE OF VOLTERRA INTEGRAL OPERATORS

H. Fujita
University of Tokyo, Japan

L. B. Rall
Mathematics Research Center
University of Wisconsin-Madison

§0. Abstract. Suppose that a given two-point boundary value problem to be solved is a perturbation of one for which a considerable amount of analytic information is available, in particular, the Green's function of the unperturbed problem is known. In this case, one can construct an analytic method for the solution of the perturbed problem which involves only the solution of a Volterra integral equation. Discretization of this method, which is a version of the classical "shooting" procedure, leads to a numerical technique for the solution of perturbed boundary value problems. Under suitable assumptions, convergence of the numerical method will be established, and estimates will be obtained for the rate of convergence. Attention will be devoted to the cases in which the unperturbed differential operator is regular or mildly singular. Linear problems will be considered first, followed by an extension of the method to the nonlinear case.

AMS(MOS) Subject Classifications - 34B05, 34B15, 45B05, 45D05, 45G05, 45G99, 45L05, 45L10, 65R05

Key Words - Boundary value problems, Fredholm integral operators, Volterra integral operators, Nonlinear integral equations, Numerical solution of integral equations

§1. Preliminaries. This section is devoted to some basic information about regular Sturm-Liouvillle boundary value problems. For more details, the books by Ince [3] or Yosida [10] may be consulted.

1.1. The regular Sturm-Liouville operator. For the sake of definiteness, attention will be confined to boundary value problems posed on the real interval $I = [0,1]$. The symbol L_0 will denote the formally self-adjoint Sturm-Liouville operator defined by

$$L_0\varphi = -\frac{d}{dx}\left(P(x)\frac{d\varphi}{dx}\right) + Q(x)\varphi \tag{1.1}$$

for all sufficiently smooth functions $\varphi = \varphi(x)$, where P,Q are smooth, real valued functions of x on $[0,1]$, and a positive constant δ exists such that

Sponsored by the United States Army under Contract No. DA-31-124-ARO-D-462.

(1.2) $$P(x) \geq \delta > 0, \quad x \in I.$$

The boundary conditions to be considered in this part of the paper are the homogeneous conditions

(1.3) $$u(0) = 0, \quad u(1) = 0,$$

or

(1.4) $$\begin{cases} u'(0)-\sigma_0 u(0) = 0, \\ u'(1)-\sigma_1 u(1) = 0, \end{cases}$$

for positive σ_0, σ_1, unless some other conditions are explicitly stated.

In order to construct the Green's function and corresponding integral operator, suppose that $\varphi = \varphi(x)$, $\psi = \psi(x)$ are functions such that

(1.5) $L_0\varphi = 0$, $\varphi(x)$ satisfies the boundary conditions at $x = 0$,

(1.6) $L_0\psi = 0$, $\psi(x)$ satisfies the boundary conditions at $x = 1$,

and their Wronskian

(1.7) $$W(\varphi,\psi) = \begin{vmatrix} \varphi(x) & \psi(x) \\ \varphi'(x) & \psi'(x) \end{vmatrix}$$

does not vanish on the interval $[0, 1]$. The existence of functions φ,ψ satisfying conditions (1.5) - (1.7) is equivalent to the existence of a bounded inverse of the differential operator L_0 subject to the given boundary conditions in some space such as $C(I)$ or $L_2(I)$, $I = [0, 1]$. Further, as

$$\frac{d}{dx} W(\varphi,\psi) = - \frac{P'(x)}{P(x)} W(\varphi,\psi),$$

it follows that

$$P(x)W(\varphi,\psi) = \text{constant}$$

for $x \in [0,1]$, and this constant value may be taken to be

$$m_0 = P(x_0)W(\varphi,\psi)_{x_0} \tag{1.8}$$

for some $x_0 \in [0, 1]$.

The functions φ,ψ may be used to construct the Green's function $G(x,y)$ for the differential operator L_0 and associated boundary conditions. One has

$$G(x,y) = \begin{cases} \frac{1}{m_0} \varphi(x)\psi(y), & x \leq y, \\ \frac{1}{m_0} \psi(x)\varphi(y), & x \geq y, \end{cases} \tag{1.9}$$

where $m_0 = P(y)W(\varphi,\psi)_y$ is independent of y. By considering the operator $\frac{1}{m_0} L_0$ instead of L_0, one can assume without loss of generality that $m_0 = 1$, and the Green's function is of the form:

$$G(x,y) = \begin{cases} \varphi(x)\psi(y), & x \leq y , \\ \psi(x)\varphi(y), & x \geq y . \end{cases} \tag{1.9'}$$

The function $G(x,y)$ is evidently continuous in the square $I \times I$. However, it is not regular at $x = y$. In $I \times I - \Delta$, where $\Delta = \{(x,x) \mid x \in I\}$ is the diagonal of $I \times I$, the smoothness of G is determined by the differentiability of φ and ψ. In particular, if $P \in C^{\alpha+1}(I)$ and $Q \in C^{\alpha}(I)$ for some $\alpha \geq 0$, then $\varphi,\psi \in C^{\alpha+2}(I)$. The following lemma will be useful later.

Lemma 1.1. Let L_0 be a regular Sturm-Liouville operator associated with certain boundary conditions. Assume that L_0^{-1} exists, where L_0^{-1} is given by the Green operator $\mathbb{G}$, which is the integral operator defined by

$$(\mathbb{G}f)(x) = \int_0^1 G(x,y)f(y)dy,$$

the kernel $G(x,y)$ being the Green's function. Moreover, assume that $P \in C^{\alpha+1}(I)$ and $Q \in C^{\alpha}(I)$ for some $\alpha \geq 0$. Then

$$\mathbb{G} = L_0^{-1} : C^{\beta}(I) \to C^{\beta+2}(I)$$

for any β such that $0 \leq \beta \leq \alpha$.

1.2. Decomposition of the operator $\mathbb{G}$. It will be shown that the Green operator $\mathbb{G}$ may be expressed as the sum of a certain Volterra operator $\mathbb{V}$, associated with L_0, and an operator of rank one. To construct the Volterra operator, consider the following initial value problem:

$$(1.10) \qquad \begin{cases} L_0 = f, \\ u(0) = u'(0) = 0 . \end{cases}$$

The solution of this problem may be expressed in the form

$$(1.11) \qquad u(x) = \int_0^x V(x,y)f(y)dy =: (\mathbb{V}f)(x) .$$

An explicit representation for $V(x,y)$ can be given if one knows two solutions φ, ψ of $L_0 u = 0$ which are independent in the sense that their Wronskian does not vanish. The standard way to do this by the variation of parameter formula [10] is to assume that $V(x,y)$ has the form

$$V(x,y) = \begin{cases} 0 , & x \leq y , \\ \alpha\varphi(x) + \beta\psi(y) = h_y(x), & y < x , \end{cases}$$

and then determine $\alpha = \alpha(y)$, $\beta = \beta(y)$ so that

$$\begin{cases} h_y(y) = 0, \\ -P(y)h_y'(y) = 1, \end{cases}$$

that is, so that $L_0 h_y = \delta(x-y)$ in terms of the Dirac δ-function [2]. This leads to the equations

$$\begin{cases} \alpha\varphi + \beta\psi = 0, \\ \alpha\varphi' + \beta\psi' = -\dfrac{1}{P}, \end{cases}$$

for α, β, and the corresponding solutions

$$\alpha(y) = \frac{\begin{vmatrix} 0 & \psi \\ -\frac{1}{P} & \psi' \end{vmatrix}}{W(\varphi,\psi)} = \frac{\psi(y)}{P(y)W(\varphi,\psi)_y} = \frac{\psi(y)}{m_0},$$

$$\beta(y) = -\frac{\varphi(y)}{m_0}.$$

This gives

$$V(x,y) = \begin{cases} 0, & x \le y, \\ \dfrac{1}{m_0}\left[\varphi(x)\psi(y) - \varphi(y)\psi(x)\right], & y < x, \end{cases} \tag{1.12}$$

where m_0 is the constant defined by (1.8). As before, the operator $\frac{1}{m_0} L_0$ could be considered in place of L_0, and one can thus assume that $m_0 = 1$ without loss of generality.

Remark 1.1. It is possible to make an alternative derivation of (1.12). Suppose that $f \in C[0,1]$ and $c \in R^1$ is a real number. It follows that $u = \mathbb{V}f + c\varphi$ is a function satisfying $L_0 u = 0$ and the boundary condition at the left endpoint $x = 0$. The method of "shooting" [4] consists of the determination of the value of c for which $\mathbb{V}f + c\varphi$ is equal to $\mathbb{G}f$. It is easy to see that the value $c = c(f)$ determined in this way is a linear functional of f which is continuous from L_2 into R^1. Thus, by the Riesz representation theorem [2, p. 20], an element $\psi_1 \in L_2$ exists such that $c(f) = (f, \psi_1)$. The relationship

(1.13) $$\mathbb{W}f + (f,\psi_1)\varphi = \mathbb{G}f$$

for all $f \in L_2$ is equivalent to

(1.14) $$V(x,y) + \varphi(x)\psi_1(y) = G(x,y)$$

for all x,y in $[0,1]$. If $x < y$, this implies that

$$\varphi(x)\psi_1(y) = G(x,y) = \varphi(x)\psi(y),$$

and thus

$$\psi_1(y) = \psi(y) .$$

Using the symmetry property of $G(x,y)$, one gets

$$V(x,y) + \varphi(x)\psi(y) = G(x,y) = \psi(x)\varphi(y), \quad x > y,$$

hence

$$V(x,y) = \psi(x)\varphi(y) - \varphi(x)\psi(y) .$$

Remark 1.2. It follows from the preceding observation that

(1.15) $$\mathbb{G}f = \mathbb{W}f + (f,\psi)\varphi ,$$

where φ,ψ are the functions appearing in the representation (1.9)' of $G(x,y)$.

Remark 1.3. Equation (1.15) shows that the Green operator $\mathbb{G}$, which is an integral operator of Fredholm type, is obtained from the Volterra operator $\mathbb{W}$ by a perturbation of rank one. In this sense, $\mathbb{W}$ is close to $\mathbb{G}$. However, $\mathbb{W}$ and $\mathbb{G}$ differ in one essential respect: $(\mathbb{I} - \lambda\mathbb{W})^{-1}$ exists as a continuous operator in $C(I)$ or $L_2(I)$ for all λ (even complex λ), whereas $(\mathbb{I}-\lambda\mathbb{G})^{-1}$

does not exist if λ is an eigenvalue of the operator $\mathbb{G}$. ($\mathbb{I}$ denotes the identity operator in the space considered.) This fact will be exploited later.

1.3. Volterra integral operators. Some well-known properties of Volterra integral operators will be summarized here for later reference. It will be supposed that the operator $\mathbb{V}$ is represented by a kernel $V = V(x, y)$ which is continuous on the triangle $\{(x, y) \mid 0 \leq y \leq x \leq 1\}$. One sets $V(x, y) = 0$ for $y > x$. Furthermore, let

$$\mu = \max_{0 \leq y \leq x \leq 1} |V(x, y)| . \tag{1.16}$$

The nth iterate $\mathbb{V}^n$ of the operator $\mathbb{V}$ will have the kernel

$$V^{(n)}(x, y) = \int_y^x V^{(n-1)}(x, t)V(t, y)dt, \quad y \leq x ,$$

with $V^{(1)}(x, y) = V(x, y)$. The following result is easily established by mathematical induction.

Lemma 1.2. For M defined by (1.16),

$$|V^{(n)}(x, y)| \leq \mu^n \frac{(x-y)^{n-1}}{(n-1)!}, \quad y \leq x , \tag{1.17}$$

$n = 1, 2, \ldots,$ where $V^{(n)}(x, y)$ is the nth iterated kernel of $V(x, y)$. In particular,

$$\|\mathbb{V}^n\| \leq \frac{\mu^n}{n!} , \tag{1.18}$$

where $\| \ \|$ denotes the operator norm on $C(I)$ or $L_2(I)$.

The iterated kernels are defined to be zero for $y > x$.

Corollary 1.2.1. The Neumann series

$$(\mathbb{I}-\mathbb{V})^{-1} = \mathbb{I} + \mathbb{V} + \mathbb{V}^2 + \cdots + \mathbb{V}^n + \cdots \tag{1.19}$$

converges in the operator norm, with

$$\|(\mathbb{I}-\mathbb{V})^{-1}\| \leq e^{\mu} . \tag{1.20}$$

Remark 1.4. If

$$|V(x,y)| \le \mu_1(x-y), \quad y \le x, \tag{1.21}$$

then

$$\begin{cases} |V^{(n)}(x,y)| \le \mu_1^n \dfrac{(x-y)^{2n-1}}{\Gamma(2n)}, & y \le x, \\ \|\mathbb{V}^n\| \le \mu_1^n \dfrac{1}{\Gamma(2n+1)}, \end{cases} \tag{1.22}$$

$n = 1, 2, \ldots$.

In this case, the estimate (1.20) can be replaced by

$$\|(\mathbb{I}-\mathbb{V})^{-1}\| \le \cosh\sqrt{\mu}\,. \tag{1.22}'$$

For Volterra operators arising from a regular Sturm-Liouville operator L_0, (1.21) and consequently (1.22) - (1.22)' hold.

1.4. Degenerate perturbation of Volterra integral operators. Consider a Volterra integral operator $\mathbb{V}$ with kernel $V(x,y)$ as defined previously, and a degenerate operator $\mathbb{S}$ of finite rank m defined by

$$\mathbb{S}u = \sum_{j=1}^{m} (u, \psi_j)\varphi_j ,$$

where φ_j, ψ_j are given linearly independent functions in $C[0,1]$. The operator

$$\mathbb{K} = \mathbb{V} + \mathbb{S}$$

is called a degenerate perturbation of the Volterra integral operator $\mathbb{V}$. The following notation will be useful: The operator

$$\mathbb{J} = (\mathbb{I} - \mathbb{V})^{-1}$$

always exists by the previous results. For φ_j, ψ_i $i,j = 1, 2, \ldots, m$ continuous, one may define the matrix $A = (a_{ij})$ by

$$a_{ij} = (\psi_i, \mathbb{J}\varphi_j) .$$

The letter d will be used to denote the determinant $d = \det(I_m - A)$, where I_m is the identity matrix of order m. The following theorem is a special case of a famous result for linear integral equations.

Theorem 1.1 (Fredholm Alternative). Consder the equation

$$u - \mathbb{K}u = f, \quad f \in X, \tag{1.23}$$

where X is either $C(I)$ or $L_2(I)$. Either (i) $d \neq 0$, and (1.23) is uniquely solvable for any $f \in X$, or (ii) $d = 0$, in which case the homogeneous equation

$$u_0 - \mathbb{K}u_0 = 0 \tag{1.24}$$

has nontrivial solutions. (That is, 1 is an element of the point spectrum $\sigma_p(\mathbb{K})$, in other words, an isolated eigenvalue of $\mathbb{K}$.)

Proof. Since $\mathbb{I} - \mathbb{V}$ and $\mathbb{J}$ are both continuous, equation (1.23) is equivalent to

$$u - \mathbb{J}\mathbb{S}u = \mathbb{J}f . \tag{1.25}$$

By definition,

$$\mathbb{J}\mathbb{S}u = \sum_{j=1}^{m} (u, \psi_j)\mathbb{J}\varphi_j .$$

Setting $\gamma_j = (u, \psi_j)$, $j = 1, 2, \ldots, m$, (1.25) becomes

$$u - \sum_{j=1}^{m} \gamma_j \mathbb{J}\varphi_i = \mathbb{J}f , \tag{1.26}$$

and since

$$(u, \psi_i) = \sum_{j=1}^{m} \gamma_j(\psi_i, \mathbb{J}\varphi_j) = (\mathbb{J}f, \psi_i) ,$$

one has

$$\gamma_i - \sum_{j=1}^{m} a_{ij}\gamma_j = (\mathbb{J}f, \psi_i) , \tag{1.27}$$

$i = 1, 2, \ldots, m$. If $d \neq 0$, then (1.27) determines the values of $\gamma_1, \gamma_2, \ldots, \gamma_m$ uniquely, and the unique solution u of (1.23) is determined correspondingly from (1.26), that is

$$u = \sum_{j=1}^{m} \gamma_j \mathbb{J}\varphi_j + \mathbb{J}f \,. \tag{1.28}$$

If $d = 0$, then the homogeneous equation

$$\gamma_i - \sum_{j=1}^{m} a_{ij}\gamma_j = 0 \tag{1.29}$$

will have nontrivial solutions belonging to some r-parameter family. Corresponding to this, (1.24) will have the r-parameter family of nontrivial solutions

$$u_0 = \sum_{j=1}^{m} \gamma_j \mathbb{J}\varphi_j \,. \tag{1.30}$$

§2. Regular Sturm-Liouville problems perturbed by lower order terms. The operator L_0 will be considered to be a regular Sturm-Liouville operator with associated boundary conditions of the type considered in §1. Here, attention will be devoted to equations of the form

$$-L_0 u + p(x)\frac{du}{dx} + q(x)u = f_1 \,, \tag{2.1}$$

subject to the boundary conditions

$$u(0) = u(1) = 0 \,, \tag{2.2}_0$$

or

$$\begin{cases} u'(0) - \sigma_0 u(0) = 0 \,, \\ u'(1) - \sigma_1 u(1) = 0 \,, \end{cases} \tag{2.2}_1$$

where $\sigma_0, \sigma_1 > 0$ are given constants. The differential operator M is defined by

$$Mu = p\frac{du}{dx} + qu \,. \tag{2.3}$$

Equation (2.1) will be dealt with from the standpoint that the operator L_0 is perturbed by the operator M of lower order.

The functions p, q entering into (2.3) are assumed to be smooth. In particular, if $P \in C^{\alpha+1}$, $Q \in C^{\alpha}$, $\alpha \geq 0$, and one assumes likewise that $p \in C^{\alpha+1}$, $q \in C^{\alpha}$, then it follows from $f_1 \in C^{\beta}$ that $u \in C^{\beta+2}$ for any β such that $0 \leq \beta \leq \alpha$, where u is any solution of (2.1).

Equation (2.1) and the corresponding boundary conditions may be transformed into equivalent integral equations by means of the Green's operator $\mathbb{G} = L_0^{-1}$ and the corresponding Volterra operator $\mathbb{V} = \mathbb{G} - (\ , \psi)\varphi$ as defined in §1. First of all,

$$u - \mathbb{G}Mu = f, \quad \text{where} \quad f = -\mathbb{G}f_1 . \tag{2.4}$$

This is equivalent to (2.1) and $(2.2)_0$ or $(2.2)_1$. Next,

$$u - \mathbb{V}Mu - (Mu, \psi)\varphi = f , \tag{2.5}$$

which is equivalent to (2.4) and hence to the original boundary value problem. By using the definitions of $\mathbb{G}$ and $\mathbb{V}$, it can be checked directly that any solution u of (2.4) or (2.5) will satisfy the boundary conditions, and also the differential equation (2.1).

Before a general treatment of equation (2.5) is given, two special cases will be considered.

Case I. $p \equiv 0$. Here, one sets

$$\begin{cases} \tilde{V}(x, y) = V(x, y)q(y) , \\ \tilde{\psi}(y) = q(y)\psi(y) , \end{cases} \tag{2.6}$$

and obtains the corresponding equation

$$u(x) - \int_0^x \tilde{V}(x, y)u(y)dy - (u, \tilde{\psi})\varphi(x) = f(x) . \tag{2.7}$$

In this case, $\tilde{V}(x, x) = 0$. One has:

Proposition 2.1. Equations (2.5) and (2.7) are equivalent.

CaseII. The boundary conditions $(2.2)_0$ hold. By integration by parts,

$$\int_0^x V(x,y)p(y)\frac{du}{dy}\,dy = V(x,y)p(y)u(y)\Big|_{y=0}^{y=x} - \int_0^x \frac{\partial}{\partial y}\{V(x,y)p(y)\}u(y)dy$$

$$= -\int_0^x \frac{\partial}{\partial y}\{V(x,y)p(y)\}u(y)dy\ ,$$

since the first term vanishes at $y = x$ because $V(x,x) = 0$, and at $y = 0$, one has $u(0) = 0$. Also,

$$\int_0^1 p(y)\frac{du}{dy}\psi(y)dy = p(y)u(y)\psi(y)\Big|_0^1 - \int_0^1 u(y)\{p(y)\psi(y)\}'dy$$

$$= -\int_0^1 u(y)\{p(y)\psi(y)\}'dy\ ,$$

since $u(0) = u(1) = 0$. Thus, for

$$(2.8)\qquad \begin{cases} \tilde{V}(x,y) = -\frac{\partial}{\partial y}\{V(x,y)p(y)\} + V(x,y)q(y)\ , \\ \tilde{\psi}(y) = -\{p(y)\psi(y)\}' + q(y)\psi(y)\ , \end{cases}$$

one obtains

$$(2.9)\qquad u(x) - \int_0^x \tilde{V}(x,y)u(y)dy - (u,\tilde{\psi})\varphi(x) = f(x)\ ,$$

an equation of the same form as (2.7). From the above derivation, one has:

Proposition 2.2. Equations (2.5) and (2.9) are equivalent.

Case III. Reduction of the general case to a system of simultaneous equations. This is similar to the standard procedure for the conversion of a differential equation of arbitrary order into a system of first order equations. One sets

$$v = \frac{du}{dx}\ ,\quad w = \begin{pmatrix} u \\ v \end{pmatrix}\ .$$

Equation (2.5) may be written in terms of u and v as

$$u(x) - \int_0^x V(x, y)\{p(y)v(y) + q(y)u(y)\}dy - (pv + qu, \psi)\varphi(x) = f(x) ,$$

since $Mu = pv + qu$. Differentiating this equation with respect to x and making use of the fact that $V(x, x) = 0$ gives

$$v(x) - \int_0^x V_1(x, y)\{p(y)v(y) + q(y)u(y)\}dy - (pv + qu, \psi)\varphi'(x) = f'(x) ,$$

where $V_1(x, y) = \frac{\partial}{\partial x} V(x, y)$.

The two equations above can be combined into a single vector equation for w. Set

$$\tilde{V}(x, y) = \begin{pmatrix} V(x, y)q(y) & V(x, y)p(y) \\ V_1(x, y)q(y) & V_1(x, y)p(y) \end{pmatrix} ,$$

$$\tilde{\psi}(y) = \begin{pmatrix} q(y)\psi(y) \\ p(y)\psi(y) \end{pmatrix} , \quad \tilde{\varphi}(x) = \begin{pmatrix} \varphi(x) \\ \varphi'(x) \end{pmatrix} , \quad \tilde{f}(x) = \begin{pmatrix} f(x) \\ f'(x) \end{pmatrix} .$$

Then

$$w(x) - \int_0^x \tilde{V}(x, y)w(y)dy - \langle w, \tilde{\psi} \rangle \tilde{\varphi}(x) = \tilde{f}(x) , \tag{2.10}$$

where $\langle w, \tilde{\psi} \rangle$ denotes the inner product

$$\langle w, \tilde{\psi} \rangle = \int_0^1 (u(y)\ v(y)) \begin{pmatrix} q(y)\psi(y) \\ p(y)\psi(y) \end{pmatrix} dy .$$

The following result is evident.

Proposition 2.3. Equations (2.5) and (2.10) are equivalent.

Case IV. Another treatment of the general case. If one sets $w = Mu$, then equation (2.5) becomes

$$u - \mathbb{V}w - (w, \psi)\varphi = f . \tag{2.11}$$

Operating on (2.11) with M gives

$$w - M\mathbb{V}w - (w, \psi)M\varphi = Mf .$$

Thus, for

$$\begin{cases} \tilde{V}(x, y) = p(x)\dfrac{\partial}{\partial x}V(x, y) + q(x)V(x, y) =: M_x V(x, y) , \\ \tilde{\varphi}(x) = M\varphi(x) = p(x)\dfrac{d\varphi}{dx} + q(x)\varphi(x) , \\ \tilde{f}(x) = Mf(x) = p(x)\dfrac{df}{dx} + q(x)f(x) , \end{cases}$$

one has the equation

$$w - \tilde{\mathbb{V}}w - (w, \psi)\tilde{\varphi} = \tilde{f} , \tag{2.12}$$

which is of the type (2.5). In order to recover u from a solution w of (2.12), it is not necessary to solve the first order differential equation $w = Mu$ for u. Instead, one has directly from equation (2.11) that

$$u = \mathbb{V}w + (w, \psi)\varphi + f , \tag{2.13}$$

and thus:

Proposition 2.4. Equation (2.5) is equivalent to the system of equations (2.12) and (2.13).

In general, then, one only has to solve an integral equation of the form

$$u - \tilde{\mathbb{V}}u - (u, \tilde{\psi})\tilde{\varphi} = \tilde{f} , \tag{2.14}$$

where $\tilde{\mathbb{V}}$ is a regular Volterra integral operator and $\tilde{\varphi}$, $\tilde{\psi}$, and $\tilde{f}$ are given smooth functions. In order to solve (2.14), put $\tilde{\mathbb{J}} = (\mathbb{I} - \tilde{\mathbb{V}})^{-1}$, which always exists. It follows that (2.14) is equivalent to

$$u - (u, \tilde{\varphi})\tilde{\mathbb{J}}\tilde{\varphi} = \tilde{\mathbb{J}}\tilde{f} , \tag{2.15}$$

or to the system

$$\begin{cases} u = \gamma\tilde{\mathbb{J}}\tilde{\varphi} + \tilde{\mathbb{J}}\tilde{f} , \\ [1 - (\tilde{\mathbb{J}}\tilde{\varphi}, \tilde{\psi})]\gamma = (\tilde{\mathbb{J}}\tilde{f}, \tilde{\psi}) , \end{cases} \tag{2.16}$$

for $\gamma = (u, \tilde{\psi})$. Thus, after $\tilde{\mathbb{J}}\tilde{\varphi}$ and $\tilde{\mathbb{J}}\tilde{f}$ have been formed, one finds the solution u of (2.15) directly from (2.16), provided that

$$1 - (\tilde{\mathbb{J}}\tilde{\varphi}, \tilde{\psi}) \neq 0 . \tag{2.17}$$

§3. Numerical solution. The solution of an integral equation of the form

$$u - \mathbb{V}u - (u, \psi)\varphi = f , \tag{3.1}$$

where $V(x, y)$, $\varphi(x)$, $\psi(x)$, and $f(x)$ are given smooth functions, may be broken down into the following steps. First, one solves the linear Volterra integral equations

$$\Phi - \mathbb{V}\Phi = \varphi, \quad F - \mathbb{V}\Gamma = f \tag{3.2}$$

for the functions

$$\Phi(x) = \mathbb{J}\varphi(x), \quad F(x) = \mathbb{J}f(x) . \tag{3.3}$$

The inner products

$$(\Phi, \psi) = \int_0^1 \Phi(x)\psi(x)dx, \quad (F, \psi) = \int_0^1 F(x)\psi(x)dx , \tag{3.4}$$

are then calculated, which leads to the scalar equation

$$[1 - (\Phi, \psi)]\gamma = (F, \psi) \tag{3.5}$$

for the constant

$$\gamma = (u, \psi) . \tag{3.6}$$

Assuming that $1 - (\Phi, \psi) \neq 0$, so that (3.5) is uniquely solvable, one obtains the solution $u(x)$ of the original equation (3.1) in the form

$$u(x) = F(x) + \frac{(F, \psi)}{1 - (\Phi, \psi)} \Phi(x) . \tag{3.7}$$

In general, one would not expect to obtain explicit representations for the functions $F(x)$, $\Phi(x)$, and the integrals appearing in (3.4). This leads to the consideration of a number of procedures for obtaining approximate solutions. For example, since the Neumann series for $\mathbb{J} = (\mathbb{I} - V)^{-1}$ is convergent, it is natural to try to obtain the values of (3.3) and (3.4) in terms of infinite series. Another approach, the one which will be considered in this paper, is to use numerical integration (Nyström's method) for the solution of the integral equations (3.2) and the evaluation of the integrals (3.4). (It is not necessary to use the same rule of numerical integration for each purpose.) Once again, the smoothness of the functions involved leads to the possibility of obtaining fairly accurate results by numerical integration.

This is in contrast to what may happen if the Nyström method is applied in a straightforward fashion to the Fredholm integral equation

$$u - \mathcal{G}u = f , \tag{3.8}$$

as the singularity in $\frac{\partial}{\partial t}G(x,t)$ at $x = t$ may affect the accuracy of a given procedure for numerical integration [1, 7].

Convergence of this method of numerical solution will be proved under fairly mild restrictions concerning the quadrature formulas to be used. It will be assumed that the interval $[0,1]$ is partitioned into N subintervals by means of the points

$$0 = \xi_0 < \xi_1 < \cdots < \xi_{N-1} < \xi_N = 1 .$$

It is not essential that these subintervals be of equal length; however, the lengths of these subintervals should go to zero uniformly as $N \to \infty$. As usual, one sets

$$f_i = f(\xi_i),\ \varphi_i = \varphi(\xi_i),\ \psi_i = \psi(\xi_i) , \tag{3.9}$$

$i = 0, 1, \ldots, N$, and one seeks approximations u_i^N to $u(\xi_i)$ by replacing (3.1) by a finite linear system of equations. This discretization is accomplished by the introduction of quadrature formulas

$$\langle u^N, \psi \rangle = \sum_{j=0}^{N}{}' C_j^N u_j^N \psi_j \tag{3.10}$$

to approximate

$$(u, \psi) = \int_0^1 u(x)\psi(x)dx ,$$

and

$$\sum_{j=0}^{i}{}' K_{ij}^N u_j^N = \sum_{j=0}^{i}{}' C_{ij}^N V(\xi_i, \xi_j) u_j^N \tag{3.11}$$

as approximations to the integrals

$$\int_0^{\xi_i} V(\xi_i, y)u(y)dy ,$$

$i = 0,1,\ldots,N$. The values of C_j^N, C_{ij}^N depend on the specific numerical integration rules used. This leads to the system

$$u_i^N - \sum_{j=0}^{i}{}' K_{ij}^N u_j^N - \langle u^N, \psi \rangle_N \varphi_i = f_i , \tag{3.12}$$

$i = 0,1,\ldots,N$, for $u^N = (u_0^N, u_1^N, \ldots, u_N^N)$. The prescription for solving this finite system is essentially the same as for the integral equation (3.1). As in [7], one solves the lower triangular systems

$$F_i^N - \sum_{j=0}^{i}{}' K_{ij}^N F_j^N = f_i, \quad \Phi_i^N - \sum_{j=0}^{i}{}' K_{ij}^N \Phi_j^N = \varphi_i , \tag{3.13}$$

$i = 0,1,\ldots,N$ for $F^N = (F_0^N, F_1^N, \ldots, F_N^N)$, $\Phi^N = (\Phi_0^N, \Phi_1^N, \ldots, \Phi_N^N)$, and forms the inner products

$$\begin{cases} \langle F^N, \psi \rangle_N = \sum_{j=0}^{N}{}' C_j^N F_j^N \psi_j , \\ \langle \Phi^N, \psi \rangle_N = \sum_{j=0}^{N}{}' C_j^N \Phi_j^N \psi_j . \end{cases} \tag{3.14}$$

Then, if

$$1 - \langle \Phi^N, \psi \rangle_N \neq 0 , \tag{3.15}$$

the solution of (3.12) is given by

$$u^N = F^N + \gamma^N \Phi^N , \tag{3.16}$$

where

$$\gamma^N = \frac{\langle F^N, \psi \rangle_N}{1 - \langle \Phi^N, \psi \rangle_N} . \tag{3.17}$$

Notice that it is not necessary to evaluate the inner product $\gamma^N = \langle u^N, \psi \rangle_N$ directly.

Ordinarily, one has $V(x, x) = 0$, which implies that $K_{ii}^N = 0$, $i = 0, 1, \ldots, N$. In this case, the amount of arithmetic required to obtain the solution (3.16) of (3.12) has been shown [7] not to exceed one division, $N^2 + 4N + 5$ multiplications, and $N^2 + 4N + 4$ additions. The following arguments may be modified easily to handle the case $K_{ii}^N \neq 0$, at the possible cost of $\frac{1}{2}(N^2 + 5N + 5)$ additional divisions. This modification will be indicated at the end of this section.

The following assumption will be made concerning the quadrature formulas used in (3.11) and (3.14): A constant α, independent of N, exists such that

$$|C_j^N| \leq \frac{\alpha}{N}, \quad |C_{ij}^N| \leq \frac{\alpha}{N}, \tag{3.18}$$

$0 \leq j \leq i \leq N$. Since it is assumed that

$$|V(x, y)| \leq \mu,$$

this in turn implies

$$|K_{ij}^N| \leq \frac{\alpha\mu}{N}, \tag{3.19}$$

$0 \leq j < i \leq N$. These inequalities may be used to obtain a uniform bound for the inverse matrices

$$J_N = (I_N - K_N)^{-1}, \tag{3.20}$$

where I_N denotes the identity matrix of order $N + 1$, and K_N is the strictly lower triangular matrix with elements

$$(K_N)_{ij} = \begin{cases} 0, & 0 \leq i \leq j \leq N, \\ K_{ij}^N, & 0 < j < i \leq N. \end{cases} \tag{3.21}$$

For this purpose, it will be convenient to introduce the norm

$$\|z\| = \max_{(i)} |z_i| \tag{3.22}$$

for vectors $z = (z_0, z_1, \ldots, z_N)$ in the space R^{N+1} of $(N+1)$-dimensional real vectors. The corresponding bound $\|A\|$ for real matrices $A = (A_{ij})$ of order $N + 1$ is

$$\|A\| = \max_{(i)} \sum_{j=0}^{N} |A_{ij}| . \tag{3.23}$$

It may be noted that the inner product $\langle , \rangle_N$ defined by (3.10) may be used to obtain the norm $\|z\| = \langle z, z \rangle_N^{\frac{1}{2}}$ on R^{N+1}, provided $C_j^N > 0$, $j = 0, 1, \ldots, N$. This norm and the corresponding matrix norm will be equivalent to (3.22) and (3.23), respectively. For computational purposes, (3.22) and (3.23) have the advantage of simplicity.

Lemma 3.1. Let $B = (B_{ij})$ be a strictly lower triangular matrix of order $N + 1$ with elements

$$B_{ij} = \begin{cases} 0 , & 0 \le i \le j \le N , \\ b_{ij}, & 0 < j < i \le N . \end{cases} \tag{3.24}$$

Furthermore, assume that a constant $\lambda > 0$ exists such that

$$|b_{ij}| \le \lambda , \tag{3.25}$$

$0 < j < i \le N$. Then, for any vector $z \in R^{N+1}$ and any positive integer p,

$$|(B^p z)_i| \le \frac{\lambda^p}{p!} i(i-1) \cdots (i-p+1) \|z\| , \tag{3.26}$$

$i = 0, 1, \ldots, N$. Thus,

$$\|B^p\| \le \frac{(\lambda N)^p}{p!} , \tag{3.27}$$

and

$$\|(I_N - B)^{-1}\| \le e^{\lambda N} . \tag{3.28}$$

Proof: Inequalities (3.26) will be established by mathematical induction on p. Note that it follows from (3.26) that

$$(B^p z)_i = 0 \tag{3.29}$$

for $0 \le i < p$. For $p = 1$,

$$(Bz)_0 = 0 ,$$

and

$$(Bz)_i = \sum_{j=0}^{i-1} b_{ij} z_j ,$$

$i = 1, 2, \ldots, N$. Thus

$$|(Bz)_i| \leq \left| \sum_{j=0}^{i-1} b_{ij} z_j \right| \leq \sum_{j=0}^{i-1} \lambda \|z\| = i\lambda \|z\| ,$$

$i = 1, 2, \ldots, N$. This verifies (3.26) for $p = 1$. Assume that (3.26) holds for $p = m$. Once again, one has

$$(B^{m+1}z)_0 = (B(B^m z))_0 = 0 .$$

Also, for $i = 1, 2, \ldots, N$,

$$(B^{m+1}z)_i = (B(B^m z))_i = \sum_{j=0}^{i-1} b_{ij}(B^m z)_j ,$$

so

$$|(B^{m+1}z)_i| \leq \sum_{j=0}^{i-1} \lambda \cdot \frac{\lambda^m}{m!} j(j-1) \cdots (j-m+1) \|z\| . \tag{3.30}$$

Actually,

$$\sum_{j=0}^{i-1} j(j-1) \cdots (j-m+1) = \sum_{j=m}^{i-1} j(j-1) \cdots (j-m+1) .$$

Another application of mathematical induction will be made to show that

$$(m+1) \sum_{j=m}^{i-1} j(j-1) \cdots (j-m+1) = i(i-1) \cdots (i-m) \tag{3.31}$$

for all positive integers $i \geq m + 1$. For $i = m + 1$, both sides of (3.31) reduce to $(m+1)!$, and thus it is valid initially. Assuming that (3.31) holds for some positive integer $i = n \geq m + 1$, one has

$$(m+1)\sum_{j=m}^{n} j(j-1)\cdots(j-m+1) = (m+1)n(n-1)\cdots(n-m+1) + (m+1)\sum_{j=m}^{n-1} j(j-1)\cdots(j-m+1)$$

$$= (m+1)n(n-1)\cdots(n-m+1) + n(n-1)\cdots(n-m)$$

$$= (n+1)n(n-1)\cdots(n-m+1)\ ,$$

which is (3.31) for $i = n + 1$. This completes the "inside" induction. From (3.30) and (3.31),

$$|(B^{m+1}z)_i| \le \frac{\lambda^{m+1}}{(m+1)!}\, i(i-1)\cdots(i-m)\,\|z\|$$

which is (3.26) for $p = m + 1$. This completes the "outside" induction and shows that (3.26) holds for all positive integers p. Inequality (3.27) follows immediately from (3.26). Since $B^{N+1} = 0$, B is nilpotent, and $(I_N - B)^{-1}$ always exists and is given by the finite Neumann series

$$(I_N - B)^{-1} = I_N + B + B^2 + \cdots + B^N\ .$$

Hence,

$$\|(I_N - B)^{-1}\| \le \sum_{p=0}^{N} \|B^p\| = \sum_{p=0}^{\infty} \|B^p\| \le \sum_{p=0}^{\infty} \frac{(\lambda N)^p}{p!} = e^{\lambda N}\ ,$$

which proves (3.28). This completes the proof of Lemma 3.1.

Lemma 3.2. For the matrix K_N defined by (3.21), one has

$$\|J_N\| = \|(I_N - K_N)^{-1}\| \le e^{\alpha\mu} \tag{3.32}$$

for all positive integers N.

Proof: This follows immediately from (3.19) and Lemma 3.1, as one may take $\lambda = \alpha\mu/N$.

Thus, it is always possible to solve equations (3.13) for F^N and Φ^N, furthermore,

$$\|F^N\| \le e^{\alpha\mu}\|f\|_\infty\ ,\quad \|\Phi^N\| \le e^{\alpha\mu}\|\varphi\|_\infty \tag{3.33}$$

uniformly in N for bounded, Riemann-integrable functions $f(x)$, $\varphi(x)$. Actually, it will not be a particular handicap to restrict the following discussion to continuous functions, and use the symbol $\| \ \|$ for the norm in R^{N+1} and $C(I)$, as its significance will always be clear from the context. In fact, some error estimates for various methods of numerical integration involve derivatives of the integrand. To use such estimates, the class of functions considered has to be restricted to those which are sufficiently smooth, and bounds for derivatives of these functions would enter into the definition of the norm.

The following assumptions will be made concerning the rates of convergence of the quadrature formulas used:

$$(3.34)\qquad \begin{cases} \left| \int_0^1 h(x)dx - \sum_{j=0}^{N}{}' C_j^N h(\xi_j) \right| \le \varepsilon_\rho(N) \|h\|_\rho , \\ \left| \int_0^{\xi_i} h(x)dx - \sum_{j=0}^{i}{}' C_{ij}^N h(\xi_j) \right| \le \varepsilon_\sigma(N) \|h\|_{\sigma,i} , \end{cases}$$

where $\varepsilon_\rho(N)$, $\varepsilon_\sigma(N)$ are called the rates of convergence of the integration formulas, with

$$(3.35)\qquad \lim_{N\to\infty} \varepsilon_\rho(N) = \lim_{N\to\infty} \varepsilon_\sigma(N) = 0 ,$$

(it is assumed that $\varepsilon_\sigma(N)$ is independent of i), and $\| \ \|_\rho$, $\| \ \|_{\sigma,i}$ are suitable norms, the latter being defined on the interval $[0, \xi_i]$.

Lemma 3.3. Suppose that $h(x)$ and $V(x,y)$ are smooth in the sense that

$$(3.36)\qquad \|V(x,\cdot)\mathbb{J}h\|_{\sigma,i} \le \mu_1(\mathbb{J}h), \qquad 0 \le x \le 1 ,$$

uniformly in i. Then,

$$(3.37)\qquad |(J_N h)_i - (\mathbb{J}h)(\xi_i)| \le \varepsilon_\sigma(N)\mu_1(\mathbb{J}h)e^{\alpha\mu} .$$

In particular, if $\{i_N\}$ is a sequence of positive integers such that

$$\lim_{N\to\infty} \xi_{i_N} = \xi ,$$

then

then

$$\lim_{N \to \infty} (J_N h)_{i_N} = (\mathbb{J}h)(\xi)$$

uniformly.

Proof: For simplicity of notation, put $g = \mathbb{J}h$. By definition,

$$(\mathbb{V}g)(\xi_i) = \int_0^{\xi_i} V(\xi_i, y)g(y)dy ,$$

and

$$(K_N g)_i = \sum_{j=0}^{i} C_{ij}^N V(\xi_i, \xi_j)g(\xi_j) .$$

Thus, from (3.34),

$$\|K_N h - \mathbb{V}g\| \leq \varepsilon_\sigma(N) \sup_{(i)} \|V(\xi_i, \cdot)g\|_{\sigma, i} \tag{3.38}$$

$$\leq \varepsilon_\sigma(N)\mu_1(g) ,$$

where the norm on the left is taken in R^{N+1}, where $Vg = (Vg(\xi_0), \ldots, Vg(\xi_N))$. One has

$$J_N h = J_N(\mathbb{I} - V)\mathbb{J}h ,$$

and

$$\mathbb{J}h = J_N(I_N - K_N)\mathbb{J}h .$$

Consequently,

$$(J_N h)_i - (\mathbb{J}h)(\xi_i) = J_N(K_N - V)\mathbb{J}h = J_N(K_N g - Vg) ,$$

from which (3.37) and the uniform convergence of the numerical values $(J_N h)_{i_N}$ to $(\mathbb{J}h)(\xi)$ follow.

Remark 3.1. Since $\mathbb{J}h$ is smoother than h, condition (3.36) is not as restrictive as it might appear.

Remark 3.2. Suppose that one chooses the simple trapezoidal rule of numerical integration [1, pp. 15-17]. Then, one may take $\|\ \|_{\sigma,i} = \|\ \|_{C^2,i}$ and

$$\varepsilon_\sigma(N) = \frac{1}{12N^2},\quad \mu_1(\mathbb{J}h) \le \kappa_1 \|\mathbb{J}h\|_{C^2} \le \kappa_2 \|h\|_C ,$$

where the constants κ_1, κ_2 depend only on $V(x, y)$. In this case,

$$\|(J_N h) - \mathbb{J}h\|_{R^{N+1}} \le \frac{\kappa_2}{12N^2} \|h\|_C . \tag{3.39}$$

Lemma 3.4. Under the above assumptions,

$$\lim_{N\to\infty} \langle J_N\varphi, \psi\rangle_N = \int_0^1 \psi(x)(\mathbb{J}\varphi)(x)dx . \tag{3.40}$$

Proof: Since φ is smooth, one may take $\mu_1(\mathbb{J}\varphi) = \kappa \|\mathbb{J}\varphi\|_\sigma$, where κ depends only on V, and $\|\ \|_\sigma$ denotes $\|\ \|_{\sigma,i}$ defined on the entire interval $[0,1]$. This gives

$$|(J_N\varphi)_i - (\mathbb{J}\varphi)(\xi_i)| \le \varepsilon_\sigma(N)\kappa e^{\alpha\mu} \|\mathbb{J}\varphi\|_\sigma ,$$

$i = 0, 1, \ldots, N$. Thus, using (3.18),

$$\left| \sum_{i=0}^N C_i^N (J_N\varphi)_i\psi_i - \sum_{i=0}^N C_i^N (\mathbb{J}\varphi)(\xi_i)\psi(\xi_i) \right| \le \tag{3.41}$$

$$\le \varepsilon_\sigma(N)\kappa e^{\alpha\mu} \|\mathbb{J}\varphi\|_\sigma \sum_{i=0}^N |C_i^N|\,|\psi_i| \le$$

$$\le \varepsilon_\sigma(N)\kappa e^{\alpha\mu} \|\mathbb{J}\varphi\|_\sigma \|\psi\|_{C^\alpha}(1 + \frac{1}{N}) .$$

Now, by (3.34),

$$\left| \sum_{i=0}^N C_i^N (\mathbb{J}\varphi)(\xi_i)\psi(\xi_i) - \int_0^1 \psi(x)(\mathbb{J}\varphi)(x)dx \right| \le \tag{3.42}$$

$$\le \varepsilon_\rho(N) \|\psi\mathbb{J}\varphi\|_\rho .$$

From (3.41) and (3.42),

$$(3.43)\qquad \left|\langle J_N\varphi,\psi\rangle_N - \int_0^1 \psi(x)(\mathbb{J}\varphi)(x)dx\right| \le \kappa_\sigma \varepsilon_\sigma(N)(1+\tfrac{1}{N}) + \kappa_\rho \varepsilon_\rho(N)\ ,$$

from which (3.40) follows immediately.

Remark 3.3. One may use the smoothness of $\mathbb{J}\varphi$ to replace $\|\mathbb{J}\varphi\|_\sigma$ by $\kappa_1\|\varphi\|_{\sigma-2}$ in (3.41), where the constant κ_1 depends only on V.

Remark 3.4. If the trapezoidal rule is used, then

$$\left|\langle J_N\varphi,\psi\rangle_N - \int_0^1 \psi(x)(\mathbb{J}\varphi)(x)dx\right| \le \frac{1}{12N^2}(\kappa_\sigma+\kappa_\rho) + \frac{1}{12N^3}\kappa_\sigma\ ,$$

where, since $\alpha = 1$,

$$\kappa_\sigma = \kappa_1 e^{\mu}\|\varphi\|_C,\quad \kappa_\rho = \kappa_1\|\varphi\|_C\|\psi\|_{C^2}\ ,$$

and κ_1 depends only on V.

Lemma 3.5. As in Lemma 3.4,

$$(3.44)\qquad \lim_{N\to\infty}\langle J_N f,\psi\rangle_N = \int_0^1 \psi(x)(\mathbb{J}f)(x)dx\ .$$

The same type of convergence rate as (3.43) may also be obtained in this case.

Theorem 3.1. Suppose that the original integral equation (3.1) is uniquely solvable for all f. Then, a positive integer N_0 exists such that for all $N \ge N_0$, the discrete system (3.12) is uniquely solvable for all f, furthermore, the solutions u^N of (3.12) converge to the solution $u(x)$ of (3.1) as $N\to\infty$.

Proof: The condition $1-(\Phi,\psi)\neq 0$ is necessary and sufficient for the unique solvability of (3.1). Consequently

$$(3.45)\qquad |1-(\Phi,\psi)| = 2\delta > 0\ .$$

Recalling that $\Phi = \mathbb{J}\varphi$, $\Phi^N = J_N\varphi$, it follows from Lemma (3.4) that a positive integer N_0 exists such that

$$|\langle \Phi^N, \psi\rangle_N - (\Phi, \psi)| \leq \delta \tag{3.46}$$

provided $N \geq N_0$. Thus, as

$$|1 - \langle \Phi^N, \psi\rangle_N| \geq \Big| |1 - (\Phi,\psi)| - |\langle \Phi^N, \psi\rangle_N - (\Phi, \psi)| \Big| \geq \delta > 0 , \tag{3.47}$$

the finite system (3.12) is uniquely solvable for all f as long as $N \geq N_0$. The convergence of the numerical solutions (3.16) to the analytic solution (3.5) follows at once from Lemmas 3.3 - 3.5.

The inequalities used in the proofs of Lemmas 3.3 - 3.5 may be used to obtain the following result.

Corollary 3.1.1. If $N \geq N_0$, then

$$|u_i^N - u(\xi_i)| \leq O(\varepsilon_\rho(N)) + O(\varepsilon_\sigma(N)) . \tag{3.48}$$

Thus, the convergence rate of the numerical solution is determined by the slower of the convergence rates of the numerical integration methods used, and is of the same order of magnitude.

The case $V(x, x) \neq 0$ will now be considered briefly. For $N \geq N_1$, where N_1 is some positive integer, condition (3.19) guarantees that

$$1 - K_{ii}^N \neq 0 , \tag{3.49}$$

in fact, for $\frac{\alpha\mu}{N} < 1$, $N \geq N_1$,

$$1 - K_{ii}^N \geq 1 - \frac{\alpha\mu}{N} , \tag{3.50}$$

$i = 0, 1, \ldots, N$. Dividing equations (3.12) by $1 - K_{ii}^N$ gives a system of the form considered above, however, the result of Lemma 3.2 should be replaced by

$$\|J_N\| \leq \frac{1}{1 - \frac{\alpha\mu}{N}} e^{\frac{\alpha\mu}{1 - \frac{\alpha\mu}{N}}} , \tag{3.51}$$

and the corresponding value used in subsequent inequalities. This modification does not alter the rates of convergence derived previously.

§4. Mildly singular Sturm-Liouville problems. The technique presented above for regular boundary value problems of Sturm-Liouville type can be extended immediately in a formal way to problems which involve various types of singularities. However, the analysis becomes more delicate, particularly for the corresponding numerical procedure. As an example, consider the problem

$$(4.1)\qquad \begin{cases} -\frac{d}{dx}\left[(1-x^2)\frac{du}{dx}\right] + q(x)u = f_1(x), \\ u(0) = 0, \\ |u(1)| < +\infty . \end{cases}$$

Assuming that $q(x)$ is "small", take

$$(4.2)\qquad L_0 u = -\frac{d}{dx}\left[(1-x^2)\frac{du}{dx}\right] .$$

The functions $\varphi(x)$ and $\psi(x)$ defined by (1.5) and (1.6), respectively, are

$$(4.3)\qquad \varphi(x) = \log\frac{1+x}{1-x}, \quad \psi(x) = 1 ,$$

up to a constant multiplier. Their Wronskian is

$$(4.4)\qquad W(x) = \begin{vmatrix} \log\frac{1+x}{1-x} & 1 \\ \frac{1}{1+x}+\frac{1}{1-x} & 0 \end{vmatrix} = -\frac{2}{1-x^2} ,$$

so that

$$(4.5)\qquad m_0 = (1-x^2)\,W(x) = -2 .$$

The corresponding Green's function (1.9) is

$$(4.6)\qquad G(x,y) = \begin{cases} -\frac{1}{2}\log\frac{1+x}{1-x}, & x \le y , \\ -\frac{1}{2}\log\frac{1+y}{1-y}, & x \ge y . \end{cases}$$

Assuming that the transformed function

$$f(x) = \int_0^1 G(x, y) f_1(y)\,dy \tag{4.7}$$

exists, the technique of §2 may be applied to obtain the perturbed Volterra integral equation

$$u(x) - \int_0^x \tilde{V}(x, y)u(y)\,dy = f(x) + (u, \tilde{\psi})\varphi(x) \tag{4.8}$$

corresponding to (2.7), where

$$\tilde{V}(x, y) = -\frac{1}{2}\left[\log\frac{1+x}{1-x} - \log\frac{1+y}{1-y}\right]q(y) = q(y)\log\sqrt{\frac{(1-x)(1+y)}{(1+x)(1-y)}}, \tag{4.9}$$

and

$$\tilde{\psi}(x) = -\frac{1}{2}q(x)\,. \tag{4.10}$$

The analysis of §1.3 does not apply directly to equation (4.8), as the kernel $\tilde{V}(x, y)$ and the right-hand side of the equation are unbounded. However, it may not be difficult to construct a theory of equations with singularities of this type. One approach would be to consider functions $v(x)$ such that

$$u(x) = \log(1 - x) \cdot v(x) \tag{4.11}$$

is bounded at $x = 1$, and derive a regular Volterra integral equation for $v(x)$. For numerical work, one would need to derive formulas for numerical integration which apply to kernels of the form (4.9) and functions (4.11) [5, pp. 237-240].

§5. Nonlinear problems. Attention will now be turned to nonlinear boundary value problems of the form

$$L_0 u(x) - F(x, u(x)) = 0\,, \tag{5.1}$$

where L_0 is a regular Sturm-Liouville operator of the form (1.1), and one has the homogeneous boundary conditions (1.3) or (1.4). Taking the Green's function $G(x, y)$ of the operator L_0 subject to the given boundary conditions to be of the form (1.9)', equation (5.1) can be transformed immediately into the nonlinear integral equation of Hammerstein type

$$u(x) - \int_0^1 G(x, y)F(y, u(y))dy = 0 , \tag{5.2}$$

or, using (1.9)',

$$u(x) - \int_0^x \psi(x)\varphi(y)F(y, u(y))dy = \varphi(x) \int_x^1 \psi(y)F(y, u(y))dy . \tag{5.3}$$

Adding $\varphi(x) \int_0^x \psi(y)F(y, u(y))dy$ to both sides of (5.3) and defining the scalar unknown c by

$$c = \int_0^1 \psi(y)F(y, u(y))dy \tag{5.4}$$

gives the new problem

$$u(x) - \int_0^x V(x, y)F(y, u(y))dy = c\varphi(x) , \tag{5.5}$$

where

$$V(x, y) = \psi(x)\varphi(y) - \varphi(x)\psi(y) . \tag{5.6}$$

Equation (5.5) thus represents a family of nonlinear Volterra integral equations depending on the parameter c. The original problem (5.1) (or the Hammerstein integral equation (5.2)) is equivalent to the system (5.4)-(5.5). If, for example, $u(x)$ satisfies (5.2), then it will satisfy (5.5) for the value of c given by (5.4). On the other hand, suppose that (5.5) has solutions $U(x;c)$ for values of the parameter c which includes a solution $c = c^*$ of the scalar equation

$$c = \Psi(c) , \tag{5.7}$$

where

$$\Psi(c) = \int_0^1 \psi(y)F(y, U(y;c))dy . \tag{5.8}$$

Then, $u(x) = U(x;c^*)$ will be a solution of (5.2) and hence of (5.1).

The structure of the problem considered here is essentially the same as in the linear case; the Hammerstein integral operator $\mathbb{H}$ defined by

$$(\mathbb{H}f)(x) = \int_0^1 G(x, y)F(y, f(y))dy \tag{5.9}$$

has been expressed as the sum of the Volterra integral operator $\mathbb{V}$ with definition

$$(\mathbb{V}f)(x) = \int_0^x V(x, y)F(y, f(y))dy , \tag{5.10}$$

and the operator Φ with one-dimensional range given by

$$(\Phi f)(x) = \varphi(x) \int_0^1 \psi(y)F(y, f(y))dy . \tag{5.11}$$

However, from an analytic and numerical standpoint, the nonlinearity of the operators $\mathbb{V}$ and Φ introduces complications which are not present in the linear case. One has to settle the problems of existence, uniqueness or multiplicity, and dependence on the parameter c of the solutions of the nonlinear integral equations (5.5), and then consider existence and uniqueness of the fixed points of the scalar operator Ψ defined by (5.8). In the linear case, the Volterra integral equations always had a unique solution which depends linearly on the parameter c, and the fixed point problem consisted of one linear equation in one unknown. To get an idea of the complexity involved in even a simple-appearing nonlinear case, one should try to solve (5.4)-(5.5) with

$$V(x, y) = x - y \tag{5.12}$$

(which corresponds to $L_0 = -\frac{d^2}{dx^2}$ with the boundary conditions (1.3)), and

$$F(y, u(y)) = u^2(y) , \tag{5.13}$$

for a solution $u(x)$ which does not vanish identically.

In the case of numerical solution of (5.4)-(5.5), additional complications arise from the fact that $U(x;c)$ and $\Psi(c)$ are only calculated approximately. It thus appears that this decomposition of the Hammerstein integral operator does not lead to a solution method which is as direct, simple, and effective as in the linear case. This situation immediately suggests the use of a linearization technique, such as Newton's method. Here, the original nonlinear problem would be approximated by a sequence of linear problems. These intermediate problems would be solved by the methods developed above for the linear case, and their solutions will converge (under certain conditions) to the solution of the nonlinear problem.

To be specific, suppose that $u_0(x)$ is a function which one may consider to be an approximate solution of (5.1), in a sense to be made more precise later, and define

$$F_0(x) = F_z(x, u_0(x)) , \tag{5.14}$$

where $F_z(x, z) = \frac{\partial}{\partial z} F(x, z)$, and

$$R_0(x) = L_0 u_0(x) - F(x, u_0(x)) . \tag{5.15}$$

To obtain the next approximation $u_1(x)$ to $u(x)$ by Newton's method, one would solve the linear problem

$$L_0 \eta_0(x) - F_0(x)\eta_0(x) = -R_0(x) \tag{5.16}$$

with the same homogeneous boundary conditions as before for

$$\eta_0(x) = u_1(x) - u_0(x) . \tag{5.17}$$

In addition to obtaining the next approximation $u_1(x)$ from the solution of (5.16), one may analyze this linear problem for satisfaction of the sufficient conditions for the convergence of the sequence of functions $\{u_n(x)\}$, obtained by Newton's method, to $u(x)$. This sequence is constructed by defining

$$\begin{cases} F_n(x) = F_z(x, u_n(x)) , \\ R_n(x) = L_0 u_n(x) - F(x, u_n(x)) , \end{cases} \tag{5.18}$$

and solving the linear problems

$$L_0\eta_n(x) - F_n(x)\eta_n(x) = -R_n(x) \tag{5.19}$$

with homogeneous boundary conditions for the functions

$$\eta_n(x) = u_{n+1}(x) - u_n(x) , \tag{5.20}$$

$n = 0, 1, 2, \ldots$. A convenient theorem to use in this connection is the one due to L. V. Kantorovič [6,9], which provides sufficient conditions for the convergence of the sequence $\{u_n(x)\}$ and error estimates [8,9] for

$$\|u(x) - u_n(x)\| = \max_{0 \le x \le 1} |u(x) - u_n(x)| \tag{5.21}$$

in the space of continuous functions on $[0,1]$. The quantities required for the application of this theorem are estimates for $\|\eta_0\|$, a bound for the inverse of the differential operator $L_0 - F_0(x)\mathbb{I}$ (or an equivalent integral operator), and a Lipschitz constant for $F_z(x, z)$ or a bound for $F_{zz}(x, z)$ in a suitably defined region in $[0,1] \times (-\infty, +\infty)$. To obtain these estimates, suppose first that L_0 has a Green's function of the form (1.9)'. The problem (5.16) may then be transformed into the perturbed Volterra integral equation

$$\eta_0(x) - \int_0^x \tilde{V}_0(x, y)\eta_0(y)dy - (\eta_0, \tilde{\psi}_0)\varphi(x) = -R_0(x) , \tag{5.22}$$

where

$$\begin{cases} \tilde{V}_0(x, y) = [\varphi(x)\psi(y) - \varphi(y)\psi(x)]F_0(x) , \\ \tilde{\psi}_0(x) = \psi(x)F_0(x) . \end{cases} \tag{5.23}$$

The method for the solution of this problem is the same as indicated in equations (3.2)-(3.7). Letting $\mathbb{V}_0$ denote the Volterra integral operator with kernel $\tilde{V}_0(x, y)$, and

$$\mathbb{J}_0 = (\mathbb{I} - \mathbb{V}_0)^{-1} , \tag{5.24}$$

then, if

$$(\mathbb{J}_0\varphi, \tilde{\psi}_0) \neq 1 , \tag{5.25}$$

one has

$$\eta_0(x) = -(\mathbb{I}_0 R_0)(x) - \frac{(\mathbb{I}_0 R_0, \tilde{\psi}_0)}{1 - (\mathbb{I}_0 \varphi, \tilde{\psi}_0)} (\mathbb{I}_0 \varphi)(x) \,. \tag{5.26}$$

Knowing $\eta_0(x)$, one can obtain $\|\eta_0\|$ directly. As an alternative estimate, (5.26) yields

$$\|\eta_0\| \le \left[1 + \frac{\|\tilde{\psi}_0\| \cdot \|\mathbb{I}_0 \varphi\|}{|1 - (\mathbb{I}_0 \varphi, \tilde{\psi}_0)|} \right] \cdot \|\mathbb{I}_0\| \cdot \|R_0\| \,. \tag{5.27}$$

If

$$|\tilde{V}_0(x, y)| \le \mu_0, \qquad 0 \le y \le x \le 1 \,, \tag{5.28}$$

then one may use (1.20) to obtain

$$\|\eta_0\| \le e^{\mu_0} \left[1 + \frac{\|\tilde{\psi}_0\| \cdot \|\mathbb{I}_0 \varphi\|}{|1 - (\mathbb{I}_0 \varphi, \tilde{\psi}_0)|} \right] \cdot \|R_0\| \,. \tag{5.29}$$

As (5.29) holds for arbitrary $R_0(x)$, one has further that

$$B_0 = e^{\mu_0} \left[1 + \frac{\|\tilde{\psi}_0\| \cdot \|\mathbb{I}_0 \varphi\|}{|1 - (\mathbb{I}_0 \varphi, \tilde{\psi}_0)|} \right] \ge \|(L_0 - F_0(x)\mathbb{I})^{-1}\| \,. \tag{5.30}$$

Now, suppose $K = K(r_0) \ge 0$ exists such that

$$\begin{cases} \|F_z(x, v(x)) - F_z(x, w(x))\| \le K \|v - w\| \,, \\ v, w \in \{u \mid \|u - u_0\| \le r_0\} \,, \end{cases} \tag{5.31}$$

or

$$\max |F_{zz}(x, z)| \le K, \ (x, z) \in [0, 1] \times [-\|u_0\| - r_0, \|u_0\| + r_0] \,. \tag{5.32}$$

(Inequality (5.32) implies (5.31); however, the Lipschitz condition (5.31) does not require $F(x, z)$ to be twice differentiable with respect to z.)

Once the above computations have been performed, one may check the satisfaction of the hypotheses of the following theorem on the convergence of Newton's method.

Theorem 5.1 (Kantorovič). If (5.25) is satisfied, and

$$h_0 = \|\eta_0\| B_0 K \leq \frac{1}{2} \tag{5.33}$$

for

$$r_0 \geq \left(\frac{1 - \sqrt{1 - 2h_0}}{h_0} \right) \|\eta_0\| , \tag{5.34}$$

then equation (5.1) has a solution $u(x)$, to which the sequence $\{u_n(x)\}$ defined by (5.18)-(5.20) converges, with [8]

$$\|u - u_n\| \leq \frac{\Theta^{2^n}}{1 - \Theta^{2^n}} \cdot \frac{2\sqrt{1 - 2h_0}}{h_0} \cdot \|\eta_0\| , \tag{5.35}$$

where

$$\Theta = \frac{1 - \sqrt{1 - 2h_0}}{1 + \sqrt{1 - 2h_0}} . \tag{5.36}$$

If one is interested also in the value of the "shooting constant" c appearing in (5.4)-(5.5), defining

$$c_n = \int_0^1 \psi(y) F(y, u_n(y)) dy , \tag{5.37}$$

one has the following result.

Corollary 5.1. Under the hypotheses of Theorem 5.1, one has

$$|c - c_n| \leq K \|\psi\| \cdot \|u - u_n\| , \tag{5.38}$$

where $\|u - u_n\|$ may be estimated by (5.35).

Remark 5.1. Using (5.29) and (5.30), inequality (5.33) may be replaced by the condition

$$\|R_0\| \le \frac{1}{2K} e^{-2\mu_0} \left[1 + \frac{\|\tilde{\psi}_0\| \cdot \|\mathbb{J}_0\varphi\|}{|1 - (\mathbb{J}_0\varphi, \tilde{\psi}_0)|} \right]^{-2} \tag{5.39}$$

on the residue function $R_0(x)$.

For numerical purposes, the methods of §3 may be applied to the linear problems (5.19) to carry out the iteration procedure. If $\mathbb{J}_0\varphi$ and $(\mathbb{J}_0\varphi, \tilde{\psi}_0)$ can be calculated accurately enough to guarantee the satisfaction of (5.25), and the other hypotheses of Theorem 5.1 can be verified, then this insures that the original problem (5.1) has a solution $u(x)$. Any result obtained numerically can be treated as an initial approximation $u_0(x)$ to $u(x)$, from which the error bound

$$\|u - u_0\| \le \frac{1 - \sqrt{1 - 2h_0}}{h_0} \|\eta_0\| \tag{5.40}$$

follows by Theorem 5.1.

REFERENCES

1. P. J. Davis and P. Rabinowitz. Numerical integration. Blaisdell, Waltham, 1967.

2. B. Friedman. Principles and techniques of applied mathematics. John Wiley & Sons, New York, 1956.

3. E. L. Ince. Ordinary differential equations. Dover, New York, 1956.

4. H. B. Keller. Numerical methods for two-point boundary-value problems. Blaisdell, Waltham, 1968.

5. I. P. Mysovskih. Lectures on numerical methods. Wolters-Noordhoff, Gronigen, 1969.

6. L. B. Rall. Computational solution of nonlinear operator equations. John Wiley & Sons, New York, 1969.

7. L. B. Rall. Numerical inversion of Green's matrices. Technical Summary Report #1149, Mathematics Research Center, University of Wisconsin-Madison, 1971.

8. L. B. Rall. Rates of convergence of Newton's method. Technical Summary Report #1224, Mathematics Research Center, University of Wisconsin-Madison, 1972.

9. L. B. Rall and R. A. Tapia. The Kantorovich theorem and error estimates for Newton's method. Technical Summary Report #1043, Mathematics Research Center, University of Wisconsin-Madison, 1970.

10. K. Yosida. Lectures on differential and integral equations. Interscience, New York, 1960.

NONLINEAR VIBRATION THEORY OF PAVEMENTS

Richard A. Weiss
Research Physicist
U. S. Army Engineer Waterways Experiment Station
Vicksburg, MS 39180

ABSTRACT. The problem of calculating the dynamic response produced by a vertical harmonic load applied to the area on the surface of a nonlinear layered elastic half space has been studied. An elastic restoring force which includes first, third, and fifth order terms in the displacement of the pavement surface is required to describe the measured nonlinear dynamic response of pavements which are subjected to dynamic loads. The equation of motion of a nonlinear oscillator is solved for this elastic restoring force, and the dynamic deflection of the pavement surface is determined as a function of the applied static and dynamic loads. The mechanical impedance of a pavement, which is modeled as a nonlinear layered elastic half space, is found to depend on the values of the static and dynamic load applied to the pavement surface.

1. INTRODUCTION. Nondestructive vibratory testing of pavements may be of importance toward predicting the performance of airfield pavements and may be used for the rapid evaluation of pavement strength. (Ref. 1-3) To be useful, the physical quantities measured by the nondestructive testing technique must be related to pavement performance. Pavement performance is measured by number of aircraft coverages on the pavement required to reach some defined condition of failure. The U. S. Army Engineer Waterways Experiment Station (WES) was requested to perform experimental and theoretical investigations to determine if the physical quantities measured by the nondestructive technique can be used for pavement evaluation and can be related to pavement performance. Some of the quantities that are measurable by the nondestructive vibratory technique are:

a. The dynamic deflection of the pavement surface versus the frequency of vibratory loading for a series of fixed static and dynamic loads.

b. The stress and strain distribution in the pavement around the vibrator measured on

instrumented pavement sections.

c. Rayleigh wave dispersion curves giving phase velocity versus wavelength measured with the wave propagation techniques.

d. The dynamic deflection of the pavement surface versus the dynamic force for a series of fixed static loads and fixed frequencies.

Most of the previous work on the nondestructive testing of pavements has treated the mechanical quantities listed in Subparagraphs a, b, and c. This paper concentrates primarily on the nonlinear response exhibited by pavements through the measurements listed in Subparagraph d. Further study of the physical quantities mentioned in Subparagraphs a, b, and c should be made in the light of the new results obtained from the study of nonlinear effects in pavements.

The overall objectives of this pavement study are:

a. The development of a mechanical model which describes the measured response of pavements to a sinusoidal dynamic loading that is applied to the pavement surface.

b. The development of a method for determining the subsurface structure of the pavement in terms of the measured dynamic response of the pavement.

The development of the pavement response model includes the following specific objectives:

a. To determine the effects of intrinsic pavement properties and structure on the dynamic load-deflection curves.

b. To determine the effects of such vibrator characteristics as dynamic load, static load, and contact area on the dynamic load deflection curves.

c. To calculate the dynamic stiffness and determine its dependence on the intrinsic properties of the pavement and on the characteristics of the vibrator used to measure this quantity.

d. To develop a theory of the nonlinear dynamic response of pavements which enables the comparison of dynamic stiffness measurements obtained from different vibrators at the same pavement location.

The theoretical work done in this paper may have applications for the nondestructive testing of roads and airport pavements. The possible practical applications of this work are twofold: (a) the use of the dynamic stiffness measurement for determining the subsurface structure of the pavement, i.e., the shear modulus and thickness for each pavement layer, and (b) the development of the capability of comparing the values of the dynamic stiffness measured by different vibrators at the same pavement location.

To achieve the objectives listed above, both theoretical and experimental studies were conducted.

The theoretical studies included:

a. The formulation of a nonlinear mechanical model to describe the response of a pavement to static and dynamic loading.

b. The determination of effects of the structure of the pavement-soil system on the parameters which appear in the nonlinear vibration model.

c. A numerical evaluation of the parameters that appear in the nonlinear model.

d. The development of formulas giving the shear modulus and layer thickness of each pavement layer in terms of quantities that are obtained directly from the measured dynamic load-deflection curves.

The experimental studies performed on both actual airport pavements and especially constructed test sections included:

a. The measurement of dynamic load-deflection curves using a vibrator developed at WES which can generate dynamic loads up to 16 kips (WES 16-kip vibrator) with a constant 16-kip static load and a constant frequency of 15 Hz.

b. The measurement of dynamic load-deflection curves at a constant static load of 16 kips for a series of fixed frequencies in the range from 10-40 Hz.

c. The measurement of dynamic load-deflection curves at a constant static load of 16 kips and a constant frequency of 15 Hz for a series of baseplates whose diameters ranged from 5-18 in.

d. The measurement of dynamic load-deflection curves at constant frequency and constant baseplate size for a range of static loads from 5-50 kips.

2. LINEAR OSCILLATOR MODEL OF PAVEMENT RESPONSE. Nondestructive vibratory testing of pavements uses a mechanical vibrator operating at a known frequency and dynamic force applied to the pavement surface to produce a time-dependent sinusoidal deflection of the pavement surface directly beneath the vibrator baseplate. The magnitude of the dynamic deflection of the pavement surface for a series of dynamic force levels and frequencies is considered to be a measure of the strength of a pavement. This section discusses a linear oscillator model used to describe the motion of the surface of a linear elastic half-space and then shows how this model fails to account for the measured values of the dynamic deflection of the pavement for a series of frequencies and dynamic loadings generated by the vibrator. The concepts of dynamic stiffness and deflection are introduced, and the separation of static and dynamic displacements is demonstrated.

The equation of motion of a mass of pavement or soil undergoing vertical oscillations on the surface of a homogeneous elastic half-space is

$$m\ddot{x} + C_H\dot{x} + k_H x = F_V(t) \tag{1}$$

where

m = lumped mass of pavement and soil

$\ddot{x}$ = acceleration of pavement surface

C_H = damping constant ($3 \cdot 4a^2\sqrt{G\gamma/g}/(1 - \nu)$, where a is the contact radius of the

vibrator baseplate, G is the shear modulus of the half-space, γ is the density by weight of the half-space, g is the acceleration due to gravity, and ν is Poisson's ratio, Reference 4)

$\dot{x}$ = velocity of pavement surface

k_H = spring constant ($4GA/(1-\nu)$, Reference 4)

x = total elastic deflection of the pavement surface under the vibrator baseplate

$F_V(t)$ = total force applied to the pavement surface (static plus dynamic)

t = time

The values of k_H and C_H are chosen to construct a damped spring model for the vertical vibrations of an elastic half-space; therefore C_H represents the radiation damping of the system. If viscous friction is present, the value of the actual damping constant may be considerably larger than the value of C_H.

The total force applied by the vibrator is written as

$$F_V(t) = F_s + F_D(t) \tag{2}$$

where F_s equals static load and $F_D(t)$ equals dynamic load. The total displacement can be written as

$$x = x_e + \xi \tag{3}$$

where x_e is the static elastic deflection of the pavement surface beneath the vibrator baseplate, F_s/k_H, and ξ is the dynamic elastic deflection of pavement surface beneath the vibrator baseplate measured from the static equilibrium deflection. Combining Equations 1, 2, and 3 gives the equation of motion as

$$m\ddot{\xi} + C_H\dot{\xi} + k_H\xi = F_D(t) \tag{4}$$

wherein all static forces and displacements have canceled. Therefore, for a linear system, the static deflection does not affect the dynamic response of the vibrator mass; only the reference point is changed by the static load.

For a sinusoidal driving force, the dynamic deflection obtained from Equation 4 is (Reference 5)

$$\xi = \frac{F_D(\omega)e^{i(\omega t-\Lambda)}}{\sqrt{\left(k_H - m\omega^2\right)^2 + C_H^2\omega^2}} \qquad (5)$$

where:

$F_D(\omega)$ = magnitude of the sinusoidal dynamic force applied to pavement surface

$e^{i(\omega t-\Lambda)}$ = complex number notation for a sinusoidal time dependence where $i = \sqrt{-1}$, ω = angular frequency, and Λ = phase angle between the dynamic load applied to the pavement surface and the dynamic deflection of the pavement surface.

Two kinds of dynamic response curves of physical interest can be obtained from Equation 5:

a. Dynamic deflection versus frequency.

b. Dynamic deflection versus dynamic force.

For a linear system, the magnitude of the maximum dynamic deflection is a simple linear function of $F_D(\omega)$ as shown in Figure 1a. The magnitude of the peak dynamic deflection as a function of frequency appears in Figure 1B for a constant force vibrator and for an eccentric mass vibrator (where the dynamic force is frequency-dependent in the manner $F_D(\omega) \sim \omega^2$). The WES 16-kip vibrator is a constant force vibrator. Therefore, for a linear system the dependence of ξ on ω is rather complicated, but the

dependence of ξ on $F_D(\omega)$ is given simply by a straight line whose slope is the dynamic stiffness. The phase angle Λ is assumed to be the same for all of the elements of the mass of pavement and subgrade which enter into motion with the vibrator mass. This is the lumped mass assumption, which requires that m be interpreted as an effective mass which vibrates in phase with the vibrator mass and has a value which is determined by requiring that the theoretical frequency response curves agree with the measured frequency response curves.

Equation 5 shows that for a linear system, the dynamic stiffness is given by

$$S = \sqrt{\left(k_H - m\omega^2\right)^2 + C_H^2\omega^2} \tag{6}$$

and depends only on the following quantities:

a. Frequency.

b. Spring constant.

c. Damping constant.

d. Lumped mass of pavement and soil.

The elastic parameters of the pavement, G and ν, and the radius of the contact area of the vibrator with the pavement enter the dynamic stiffness through k_H and C_H as seen from Equation 6 and the expressions for the spring constant and damping constant. For a linear oscillator model, the dynamic stiffness does not depend on the dynamic load or on the equilibrium elastic deflection, i.e., ξ is a linear function of $F_D(\omega)$. However, the experimental values of the dynamic stiffness of pavements indicate a strong dependence of the dynamic stiffness on the dynamic load and on the static elastic equilibrium displacement of the pavement surface. Therefore, the linear oscillator model is insufficient to describe the response of pavements to dynamic loadings, and a nonlinear oscillator model is required to explain the experimental data.

3. THE NONLINEAR MECHANICAL MODEL. In this section, a nonlinear mechanical model is developed to describe the response of a pavement-subgrade system to a sinusoidal dynamic loading applied to the surface of the pavement by a vibrator. The model is developed in three basic steps:

a. The determination of the nonlinear pavement-restoring force in terms of three parameters: the linear elastic parameter of a nonlinear pavement, the third-order nonlinear elastic parameter, and the fifth-order nonlinear elastic parameter.

b. The solution of the motion equation (4) for the case of the nonlinear pavement-restoring force and the subsequent calculation of the dynamic stiffness and deflection of the pavement as a function of the static and dynamic loads exerted by the vibrator.

c. The determination of the parameters k_{00}, b, and e in terms of the elastic constants of the layered pavement-subgrade system and in terms of the finite depth of influence that a static surface load produces in this system.

If for a fixed frequency the dynamic deflection of the pavement surface is not directly proportional to the dynamic force, the system is said to be nonlinear. The experimental data indicate that this is the case for most asphaltic concrete (AC) pavements and for some portland cement concrete (PCC) pavements. It will be shown that the nonlinear behavior of a pavement undergoing forced sinusoidal vibrations can produce very different values of dynamic stiffness such as those measured at the same location by different mechanical vibrators. Therefore, it is important to be able to account for the nonlinear effects by a simple physical model.

A physical and mathematical model for the nonlinear response of pavements can be derived which will account for the dependence of the impedance values of the type of vibrator used to determine then, i.e., on the static weight, dynamic load, and contact area of the vibrator. This paper will show that it is possible to describe the dependence of the measured values of pavement dynamic stiffness on the

physical characteristics of the vibrator by introducing three parameters to describe the nonlinear pavement-restoring force.

The pavement-restoring force is the force that the bulk pavement exerts on the lumped pavement mass from below. In linear Equation 1, the pavement-restoring force is simply $k_H x$. In general, the pavement restoring force is not equal to the force generated by the vibrator; only for the static case are these two forces equal. The first task to be accomplished is the development of a mathematical expression for the pavement-restoring force which satisfies the following two very general criteria:

a. The mathematical form of the pavement-restoring force will be sufficiently general so that the nonlinear dynamic response of the pavement that is calculated from this restoring force will be adequate to describe the experimental nonlinear dynamic load-deflection curves.

b. Only terms based on sound physical theory are included in the mathematical form of the pavement-restoring force.

The form of the nonlinear elastic pavement-restoring force used in the nonlinear model is determined by requiring the restoring force to be antisymmetric in the deflection of the pavement surface, i.e.,

$$F_P(x) = -F_P(-x) \tag{7}$$

where $F_P(x)$ equals the pavement-restoring force. Equation 7 is satisfied for the linear case, $F_P = k_H x$. A simple nonlinear pavement-restoring force which satisfies Equation 7 and which is found to be adequate to describe the dynamic load-deflection curves for pavements, is

$$F_P(x) = k_{00}x + bx^3 + ex^5 \tag{8}$$

where k_{00} equals the linear elastic parameter of a nonlinear pavement while b and e equal respectively the third- and fifth-order nonlinear parameters. The experimental data

indicate that at least two nonlinear elastic parameters, b and e , are required to describe AC and PCC pavements. The linear spring constant k_{00} which appears in Equation 8 is not in general equal to the spring constant k_H which describes the homogeneous linear elastic half-space.

The equation of motion of the oscillating lumped pavement and soil mass can now be written using the expression for the nonlinear pavement-restoring force derived in the previous section. This equation of motion cannot be completely separated into static and dynamic parts as was the case for the linear elastic system.

The equation of motion for the nonlinear spring is given by

$$m\ddot{x} + C\dot{x} + k_{00}x + bx^3 + ex^5 = F_V(t) \tag{9}$$

where C is the damping constant of the pavement-vibrator system, and m is the in-phase lumped mass of the pavement and subgrade. The value of C is larger than the value of the radiation damping constant C_H which appears in Equation 6 because C describes several material damping processes in addition to the dissipation of energy by mechanical radiation. Equation 9 can be greatly simplified by choosing a new origin of coordinates as in Equation 3, such that the motion is described in terms of coordinates measured relative to the static equilibrium deflection. By itself the static load produces a static deflection given by

$$F_s = k_{00}x_e + bx_e^3 + ex_e^5 \tag{10}$$

Substituting Equation 10 into Equation 9 enables the equation of motion to be written as

$$m\ddot{x} + C\dot{x} + k_{00}(x - x_e) + b\left(x^3 - x_e^3\right) + e\left(x^5 - x_e^5\right) = F_D(t) \tag{11}$$

Using Equation 3 and the following algebraic identities:

$$x^3 - x_e^3 = (x - x_e)\left(x^2 + xx_e + x_e^2\right) \tag{12}$$

$$x^5 - x_e^5 = (x - x_e)\left(x^4 + x^3x_e + x^2x_e^2 + xx_e^3 + x_e^4\right) \tag{13}$$

allows Equation 11 to be rewritten as

$$m\ddot{\xi} + C\dot{\xi} + k_0\xi + b\xi^3 + e\xi^5 = F_D(t) \tag{14}$$

where k_0 is the effective quasi-static spring constant and is defined by

$$k_0 = k_{00} + 3bx_e^2 + 5ex_e^4 + g(x_e\xi) \tag{15}$$

and

$$g(x_e\xi) = 3bx_e\xi + 10ex_e^3\xi + 10ex_e^2\xi^2 + 5ex_e\xi^3 \tag{16}$$

Equation 14 is a generalization of the Duffing Equation (Reference 6).

In this section, the equation which determines the amplitude of the sinusoidal dynamic deflection of the pavement surface beneath the vibrator mass is developed. The amplitude equation is expressed in terms of an effective spring constant which in turn depends on the static and dynamic deflections of the pavement surface. The dynamic stiffness for the nonlinear system will eventually be expressed in terms of this effective spring constant.

The functions $k_0(x_e\xi)$ and $g(x_e\xi)$ are time-dependent, and therefore Equation 14 is very difficult to solve exactly. Under special conditions to be described, the coefficient which appears in Equation 14 may be taken to be independent of time, thereby making this equation somewhat easier to solve. For harmonic motion, the dynamic force applied to the pavement surface by the vibrator can be written as

$$F_D(t) = F_D(\omega)e^{i\omega t} \tag{17}$$

$$\xi(t) = Ae^{i(\omega t-\Lambda)} \tag{18}$$

where A equals the amplitude of the dynamic deflection of the pavement surface directly beneath the vibrator base-plate. The dynamic deflection of the lumped mass is assumed to be equal to the dynamic deflection of the pavement surface. For the case in which the dynamic deflection amplitude is much less than the static equilibrium deflection, $A << x_e$,

$g(x_e, \xi) \approx 0$ can be used in Equation 15, while for the case where $A \approx x_e$, the time-averaged value of $g(x_e, \xi) \approx 10ex_e^4$ can be used in Equation 13. For the two special cases, the coefficient k_0 can be written as

$$k_0 = k_{00} + 3bx_e^2 + 5ex_e^4 \qquad A << x_e \tag{19}$$

$$k_0 = k_{00} + 3bx_e^2 + 15ex_e^4 \qquad A \approx x_e \tag{20}$$

A simple linear interpolation formula for k_0 is given by

$$k_0 = k_{00} + 3bx_e^2 + 5\left(1 + 2\frac{A}{x_e}\right)ex_e^4 \tag{21}$$

It should be noted that the choice of k_0 as time-independent is an approximation which becomes invlaid for large dynamic deflections.

Even with the coefficient k_0 assumed to be time-independent, Equation 14 is a nonlinear equation. However, it can be shown that Equation 14 can be cast into the form of an equivalent linear system whose amplitude equation is (Reference 6)

$$A^2\left[\left(k - m\omega^2\right)^2 + C^2\omega^2\right] = F_D^2(\omega) \tag{22}$$

provided an effective spring constant is introduced which is defined by

$$k = k_0 + \mu bA^2 + \eta eA^4 \tag{23}$$

where k_0 is given by Equation 21, $\mu = 3/4$, and $\eta = 5/8$. The effective spring constant, k, is seen to be a function of the amplitude of the dynamic deflection and also depends on the static equilibrium deflection through the coefficient

k_0 . If the static load applied to the pavement by the vibrator were zero, then $x_e = 0$, $g(x_e \xi) = 0$, and $k_0 = k_{00}$. For this case, there could be no coupling of terms between x_e and ξ and the effective spring constant would depend only on the amplitude of the dynamic deflection. On the other hand, if the dynamic load were zero, the effective spring constant would be $k = k_0$ and would depend only on the static equilibrium deflection.

4. CALCULATION OF THE DYNAMIC STIFFNESS AND THE DEFLECTION AMPLITUDE FOR NONLINEAR PAVEMENTS. This section considers the calculation of the dynamic stiffness and the dynamic deflection of the pavement surface and dynamic forces generated by the vibrator. The deflection amplitude equation (22) derived in the previous section is expanded in powers of the deflection amplitude, A , to give a tenth-order algebraic equation for the determination of A . Infinite series expansions for the dynamic amplitude and the dynamic stiffness are obtained as solutions to this equation. These solutions express the dynamic stiffness and deflection as functions of the dynamic load generated by the vibrator and the static deflection of the pavement surface. The static deflection is then expressed in terms of the static load, so that finally the dynamic stiffness and deflection are expressed in terms of the static and dynamic loads at which the vibrator is operated.

The explicit equation for the dynamic deflection amplitude will now be calculated. The dynamic stiffness of the pavement which is described by a nonlinear oscillator is given by

$$S^2 = \left(k - m\omega^2\right)^2 + C^2\omega^2 \tag{24}$$

where the effective spring constant is given by Equation 23. The dynamic stiffness depends on the amplitude of the dynamic displacement and the static equilibrium deflection. The amplitude of the dynamic deflection is determined by Equation 22, which may be written as

$$A^2S^2 = F_D^2(\omega) \tag{25}$$

Using Equations 23 and 24, the amplitude equation (25) can be written as

$$(26) \qquad S_0^2 A^2 + 2\mu b\left(k_0 - m\omega^2\right)A^4 + 2\eta\left(k_0 - m\omega^2\right)e + \mu^2 b^2 A^6 + 2\mu\eta be A^8 + \eta^2 e^2 A^{10} = F_D^2(\omega)$$

where S_0 is the value of the dynamic stiffness obtained from S by taking $k = k_0$ (or equivalently $A = 0$ in Equation 23) and is defined by the equation:

$$(27) \qquad S_0^2 = \left(k_0 - m\omega^2\right)^2 + C^2\omega^2$$

Whereas the simple linear system produces a linear equation for the calculation of the dynamic displacement in terms of the dynamic force, the nonlinear system appropriate to describe dynamic pavement response produces a fifth-order equation for calculating A^2 in terms of $F_D(\omega)$. The value S_0 appearing in Equation 27 is the dynamic stiffness in the limit of zero dynamic loading.

The tenth-order equation (26) will now be solved for the dynamic amplitude A which will take the form of an infinite series expansion. The dynamic stiffness is calculated in terms of A by Equation 25 so that S also will have the form of an infinite series expansion.

The solution of Equation 26 for the amplitude of the dynamic displacement in terms of the amplitude of the dynamic force is, in general, difficult to obtain analytically. For the case in which the dynamic force is not very large, the amplitude of motion and the dynamic stiffness are easily obtained from Equation 26 in the form of

$$(28) \qquad A = \frac{F_D(\omega)}{S_0}\left(1 + \alpha_1\psi + \alpha_2\psi^2 + \cdots\right)$$

(29) $$S = S_0\left(1 + \beta_1\psi + \beta_2\psi^2 + \cdots\right)$$

where

α_1, α_2 = coefficient appearing in the power series expansion of the amplitude of the dynamic deflection

ψ = expansion parameter

β_1, β_2 = coefficients appearing in the power series expansion of the dynamic stiffness

The values of ψ , α_1 , α_2 , β_1 , and β_2 can be obtained by combining Equations 26 and 28 with the following results:

(30) $$\psi = \frac{F_D^2(\omega)}{S_0^4}$$

and

(31) $$\alpha_1 = -ab\left(k_0 - m\omega^2\right)$$

(32) $$\alpha_2 = \frac{7}{2}a^2b^2\left(k_0 - m\omega^2\right)^2 - S_0^2\left[e\left(k_0 - m\omega^2\right) + \frac{a^2b^2}{2}\right]$$

(33) $$\beta_1 = ab\left(k_0 - m\omega^2\right)$$

(34) $$\beta_2 = S_0^2\left[ne\left(k_0 - m\omega^2\right) + \frac{a^2b^2}{2}\right] - \frac{5}{2}a^2b^2\left(k_0 - m\omega^2\right)^2$$

The solutions given in Equations 28 and 29 are valid provided the dynamic load is not so large as to prevent the convergence of these power series solutions. Equations 28-29 have been derived from Equation 25 and give the fundamental description

of the nonlinear dynamic load-deflection curves. These equations will be fitted to experimental dynamic load-deflection curves.

Equation 28 shows that the amplitude of the dynamic deflection is not a linear function of F_D but approaches the linear condition for small F_D or large S_0. The linear system can be regained by setting $b = e = 0$ in Equations 28-34. When $F_D = 0$, $A = 0$ and $S = S_0$, and the dynamic stiffness depends only on the static equilibrium deflection. If $F_s = 0$, the deflection and stiffness are given by Equations 28-34 with the provision that k_0 be replaced by the constant k_{00}. The static load through Equation 10. Therefore, in general, the dynamic stiffness of a nonlinear system will depend on the magnitude of F_D and F_s. The dependence of S on F_D enters through the expansion parameter ψ given in Equation 30, while the dependence of S on F_s enters through the function S_0 given by Equation 27.

The expression for the dynamic amplitude A given in Equation 28 shows that A does not depend linearly on F_D. The departure from linearity is due to the terms $\alpha_1\psi$, $\alpha_2\psi^2$, ..., that appear in Equation 28. It is desirable to determine the physical quantities which determine the degree of departure from the linear condition, $A = F_D/S_0$. In the range of small F_D, the predominant term describing the nonlinear behavior of the deflection of the pavement mass is obtained from Equations 28, 30, and 31 as follows:

$$\alpha_1\psi = -ub\left(k_0 - m\omega^2\right)\frac{F_D^2}{4S_0} \tag{35}$$

In general the degree of nonlinearity depends on four quantities:

a. The magnitude of the nonlinear parameters b and e .

b. The relative magnitudes of F_S and F_D .

c. The frequency at which the vibrator is operated.

d. The static stiffness S_0 of the pavement-vibrator system.

The parameter ψ , which appears in the infinite series expansion for A and S in Equations 28-30, depends inversely on S_0 in the form S_0^{-4} , and therefore it follows that the dynamic load-deflection curves of stiff pavements are more linear than those of the more flexible pavements. Thus concrete pavements are expected to have a more linear response to a dynamic loading than do the more flexible asphalt pavements. The value of S_0 includes the effects of the subgrade as well as the effects of each layer in the pavement.

It is clear from Equations 28 and 35 that the degree of departure from the linear condition expected for the response of a pavement to an applied vibratory load at the pavement surface depends on the frequency at which the vibrator is operated. In particular, it is apparent from Equation 35 that the first-order nonlinear term is frequency-dependent and that this first-order term will vanish at a special critical frequency for which $k_0 - m\omega^2 = 0$.

It is a characteristic property of the first-order nonlinear term (Equation 35) that there is a critical frequency for which this term vanishes; the critical frequency is defined by

$$\omega_c^2 = \frac{k_0}{m} \tag{36}$$

where ω_c is the critical angular frequency. In terms of

the critical frequency, the first-order nonlinear coefficient can be written as

$$\alpha_1 = -\mu b m \left| \omega_c^2 - \omega^2 \right| \tag{37}$$

At the critical frequency, the departure from a linear system occurs only through the second-order and higher terms in ψ, i.e., $\alpha_2 \psi^2 + \alpha_3 \psi^3 + \ldots$. Therefore at the critical frequency, the pavement response for small dynamic loads should be nearly linear. The critical frequency depends on the vibrator characteristics as well as on pavement properties. The connection between the resonance frequency and the critical frequency is obtained from Equation 23 as follows:

$$\omega_R^2 = \omega_c^2 + \frac{\mu b A^2 + \eta e A^4}{m} - 2\left(\frac{C}{2m}\right)^2 \tag{38}$$

where ω_R is the resonance angular frequency. In general $\omega_R < \omega_c$.

In addition to a critical frequency, $\omega_c(F_s)$ which depends on the static load of the vibrator, there is a critical static load F_{sc} for each operating frequency of the vibrator, which is defined from Equation 35 by $\alpha_1 = 0$ or

$$k_0(F_{sc}) = m\omega^2 \tag{39}$$

Using Equations 21, 37, and 39, the first-order coefficient α_1 can be written as

$$\alpha_1 = \mu b \left[k_0(F_s) - k_0(F_{sc}) \right] \tag{40}$$

$$\alpha_1 = + \frac{3\mu b^2}{k_{00}^2} \left(F_{sc}^2 - F_s^2 \right) + \cdots \tag{41}$$

It is possible to operate a vibrator at the critical condition by adjusting either the frequency or the static load of the vibrator.

An approximately linear dynamic deflection versus dynamic force curve occurs at the critical frequency. For an arbitrary frequency, the departure from this approximately linear curve is positive or negative depending on the sign of the parameter α_1 in Equation 28 and 37. The sign of the parameter α_1 depends on the sign of the parameter b and whether $\omega \gtrless \omega_c$ or $F_s \gtrless F_{sc}$. It can be shown that b is generally negative for pavements. For the case $b < 0$, which corresponds to the case in which the shear modulus of the half-space is constant with depth, or to the case of a layered system which has G decreasing with depth, as is usually the case with pavements, the dynamic stiffness and deflection versus dynamic force curves are shown schematically in Figure 2a. For the case $b > 0$, which may be possible for the situation in which the shear modulus of a layered system increases rapidly with depth, the dynamic stiffness and displacement versus dynamic force curves are shown schematically in Figure 2b. Therefore the choice of sign $b < 0$ has physical relevance to pavement problems. For the choice $b < 0$, the sign of the parameter α_1 can be positive or negative depending on whether $\omega < \omega_c$ or $\omega > \omega_c$, respectively. The value of ω_c can be determined by observing the frequency which produces the most linear load-deflection curve. The algebraic signs of b and e can be determined from the manner in which the dynamic load-deflection curves bend away (as in Figures 2a and 2b) from the approximately linear load-deflection curve which occurs at $\omega = \omega_c$.

It should be pointed out that the definition of dynamic stiffness is the ratio of the dynamic load to the dynamic deflection for each point on the dynamic load-deflection curve.

The explicit dependence of the dynamic stiffness on F_D is given by Equations 28-34. These equations will also give the explicit dependence of S on F_s, provided that the static

equilibrium displacement x_e is expressed explicity as a function of F_s by using Equation 10. Because Equation 10 is an equation of fifth degree, numerical methods are generally required for its solution. However, in the extreme of very large and very small values of F_s, analytical solutions of this equation are possible. For a very small static load, the equilibrium elastic displacement is given by

$$x_e = \frac{F_s}{k_{00}} \tag{42}$$

For somewhat larger values of F_s, the cubic term manifests itself and x_e may be obtained from the approximate equation:

$$x_e^3 + \frac{k_{00}}{b} x_e - \frac{F_s}{b} = 0 \tag{43}$$

The discriminant of this cubic equation is

$$D = \frac{F_s^2}{4b^2} + \frac{k_{00}^3}{27b^3} \tag{44}$$

and is negative for small F_s when $b < 0$. For the condition $D < 0$, the solution of the cubic equation can be written as (Reference 7)

$$x_e = \sqrt{-\frac{4k_{00}}{3b}} \cos\left(\frac{\phi}{3}\right) \tag{45}$$

$$\cos \phi = -\frac{F_s}{\sqrt{-\frac{4k_{00}^3}{27b}}} \tag{46}$$

where ϕ is the angle which appears in the solution of this cubic equation. In the limit of small F_s (or small b), the cosine term in Equation 45 has the value

$$(47) \qquad \cos\left(\frac{\phi}{3}\right) = \frac{F_s}{\sqrt{-\frac{4k_{00}^3}{3b}}} + \frac{F_s^3}{\left(-\frac{4k_{00}^3}{3b}\right)^{3/2}}$$

Combining Equations 45 and 47 gives

$$(48) \qquad x_e = \frac{F_s}{k_{00}} - \frac{bF_s^3}{k_{00}^4}$$

The solutions of Equations 45 and 48 have been derived for $b < 0$ and are therefore applicable to pavements. It can be shown that Equation 48 is also valid for $b > 0$.

With increasing values of F_s, the fifth-order terms become dominant, and in this region the approximate solution for x_e is

$$(49) \qquad x_e = \left[\frac{F_s}{e} - \frac{k_{00}}{e}\left(\frac{F_s}{e}\right)^{1/5} - \frac{b}{e}\left(\frac{F_s}{e}\right)^{3/5}\right]^{1/5}$$

Equation 10 is easily solved for the general case of an arbitrary value of F_s by using a digital computer. A schematic graph of x_e versus F_s for pavements ($b < 0$) is given in Figure 3a while the corresponding graph for a $b > 0$ formation is given in Figure 3b.

The spring constant k_0 given by Equations 19-21 has a conventional interpretation only when the dynamic deflection amplitude satisfies $A \ll x_e$. For $A \neq 0$ the spring

constant has the approximate value given by Equation 21. Using Equations 19, 48, and 49, the value of k_0 for zero dynamic amplitude and for small F_s is

$$(50) \qquad k_0 = k_{00} + 3b\left(\frac{F_s}{k_{00}} - \frac{bF_s^3}{k_{00}^4}\right)^2 + 5e\left(\frac{F_s}{k_{00}} - \frac{bF_s^3}{k_{00}^4}\right)^4$$

while for large F_s

$$(51) \qquad k_0 = k_{00} + 3b\left[\frac{F_s}{e} - \frac{b}{e}\left(\frac{F_s}{e}\right)^{3/5} - \frac{k_{00}}{e}\left(\frac{F_s}{e}\right)^{1/5}\right]^{2/5}$$

$$+ 5e\left[\frac{F_s}{e} \quad \frac{b}{e}\left(\frac{F_s}{e}\right)^{3/5} \quad \frac{k_{00}}{e}\left(\frac{F_s}{e}\right)^{1/5}\right]^{-4/5}$$

For very small F_s, Equation 50 can be rewritten as

$$(52) \qquad k_0 = k_{00} + 3b\left(\frac{F_s}{k_{00}}\right)^2 + \left(5e - 6\frac{b^2}{k_{00}}\right)\left(\frac{F_s}{k_{00}}\right)^4$$

while for very large F_s, Equation 51 can be rewritten as

$$(53) \qquad k_0 = 5e\left(\frac{F_s}{e}\right)^{4/5}$$

Equations 10 and 21 are easily solved simultaneously on a digital computer to give the general solution, $k_0 = k_0(F_s)$. Figures 4a and b show respectively the dependence of k_0 on

F_s for pavements ($b < 0$) and for $b > 0$ formations. For the case $b < 0$, the function k_0 exhibits a local minimum value for some value of F_s (or x_e).

The dynamic stiffness defined by Equation 24 depends on the damping constant C of the pavement as well as on the effective spring constant k. This damping constant is not in general equal to the damping constant for the homogeneous linear elastic half-space, C_H, that was defined in the damping constant expression. A theoretical calculation of C was not made in this paper; however, the nonlinear elastic nature of flexible pavements gives rise to a simple method of estimating the value of the damping constant. Equations 23, 34, and 36 show that when $\omega = \omega_c$, the dynamic stiffness has the critical value:

$$(54) \qquad S_c = \sqrt{\left(\mu b A^2 + \eta e A^4\right)^2 + C^2\omega_c^2}$$

$$\approx \sqrt{2C\omega_c}$$

Therefore a measure of the damping constant can be determined directly from the dynamic load versus deflection curves by measuring the critical value of the dynamic stiffness. An approximation to the value of the damping constant is thus given by

$$(55) \qquad C = \frac{S_c}{\sqrt{2\omega_c}}$$

5. SUMMARY. A nonlinear mechanical model describing the dynamic properties of a pavement-vibrator system has been developed which describes the nonlinear dynamic response of a pavement to a sinusoidal loading applied to the pavement

surface. Theoretical expressions are developed for the dynamic stiffness of a pavement measured by a mechanical vibrator which are expressed in terms of the static load, dynamic load, and vibrator baseplate size and in terms of the linear and nonlinear pavement parameters. The nonlinear mechanical model gives an analytical correlation among the values of the dynamic stiffness measured by different vibrators at the same pavement location. Experimental tests were done to determine the validity of the theoretical pavement response model. The experimental and theoretical results are in good agreement (Reference 8).

The nonlinear mechanical model developed in this paper gives the following conclusions:

a. Third- and fifth-order nonlinear terms in the displacement are required to describe the dynamic load-deflection response of actual pavement systems.

b. The theoretical nonlinear oscillator model of pavement response to a dynamic loading shows that stiff pavements have a more linear dynamic load-deflection curve than flexible pavements.

c. At specific critical frequencies the dynamic load-deflection curves become essentially linear at low values of the dynamic force.

REFERENCES

1. Hall, J. W., JR., "Nondestructive Testing of Pavements: Tests on Multiple-Wheel Heavy Gear Load Sections at Eglin and Hurlburt Airfields," Technical Report No. AFWL-TR-71-64, Mar 1972, Air Force Weapons Laboratory, Kirtland Air Force Base, N. Mex.

2. _____, "Nondestructive Testing of Pavements: Final Test Results and Evaluation Procedure," Technical Report No. AFWL-TR-72-151, Jun 1973, Air Force Weapons Laboratory, Kirtland Air Force Base, N. Mex.

3. Balakrishna Rao, H. A., "Nondestructive Evaluation of Airfield Pavements (Phase I)," Technical Report No. AFWL-TR-71-75, Dec 1971, Air Force Weapons Laboratory, Kirtland Air Force Base, N. Mex.

4. Lysmer, J., "Vertical Motion of Rigid Footings," Contract Report No. 3-115, Jun 1965, U. S. Army Engineer Waterways Experiment Station, CE, Vicksburg, Miss.; prepared by University of Michigan under Contract No. DA-22-079-eng-340.

5. Rocard, Y., General Dynamics of Vibrations, Ungar Publishing Co., New York, 1960.

6. McLachlan, N. W., Theory of Vibrations, Dover Publishing, New York, 1951.

7. Turnbull, H. W., Theory of Equations, Interscience, New York, 1957.

8. Weiss, R. A., "Nondestructive Vibratory Testing of Airport Pavements, Volume II: Theoretical Study of the Dynamic Stiffness and Its Application to the Vibratory Nondestructive Method of Testing Pavements," Technical Report FAA-RD-73-205-II, April 1975, Department of Transportation, Federal Aviation Administration, Systems Research and Development Service, Washington, D. C.

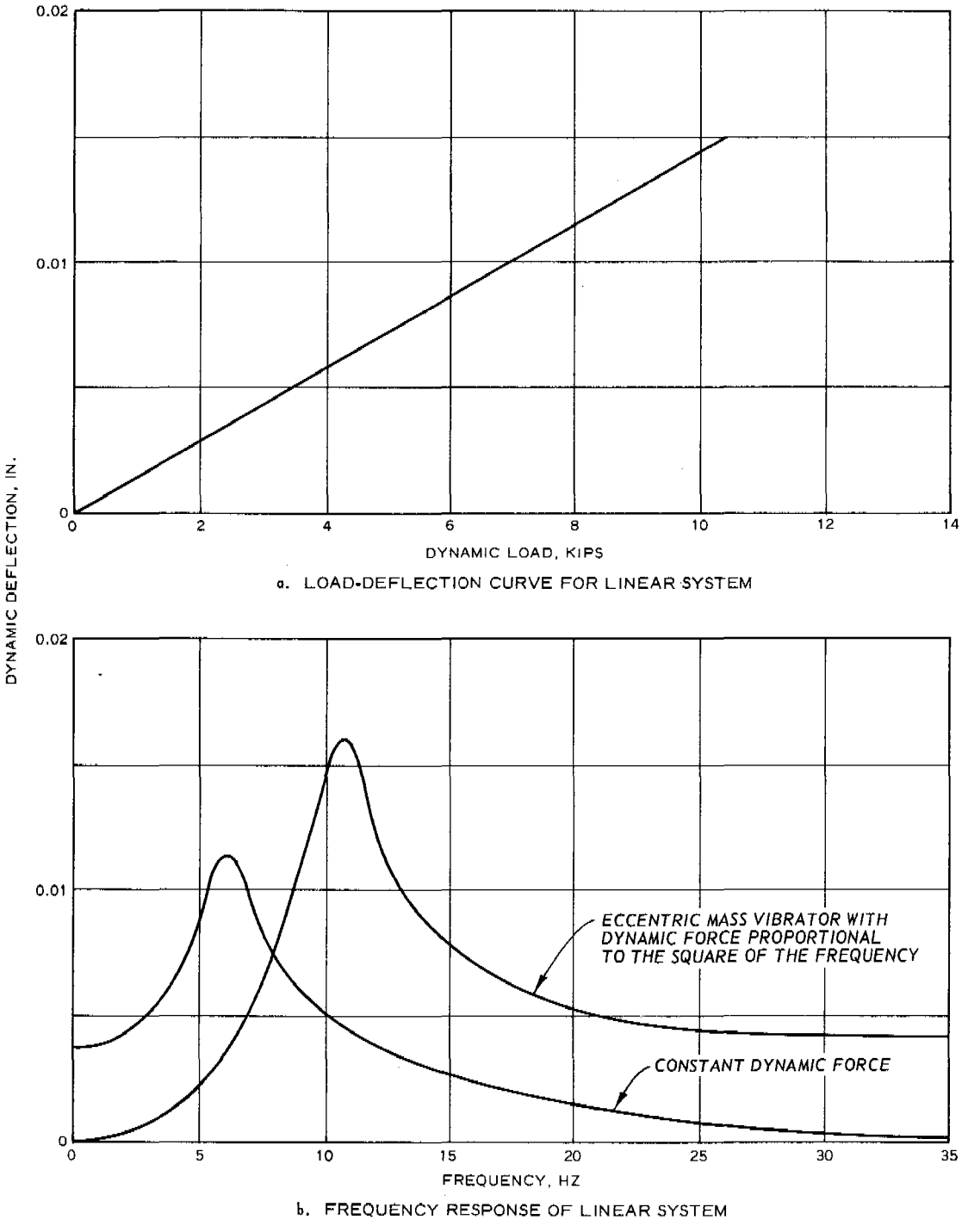


Figure 1. Typical dynamic response of the linear spring model.

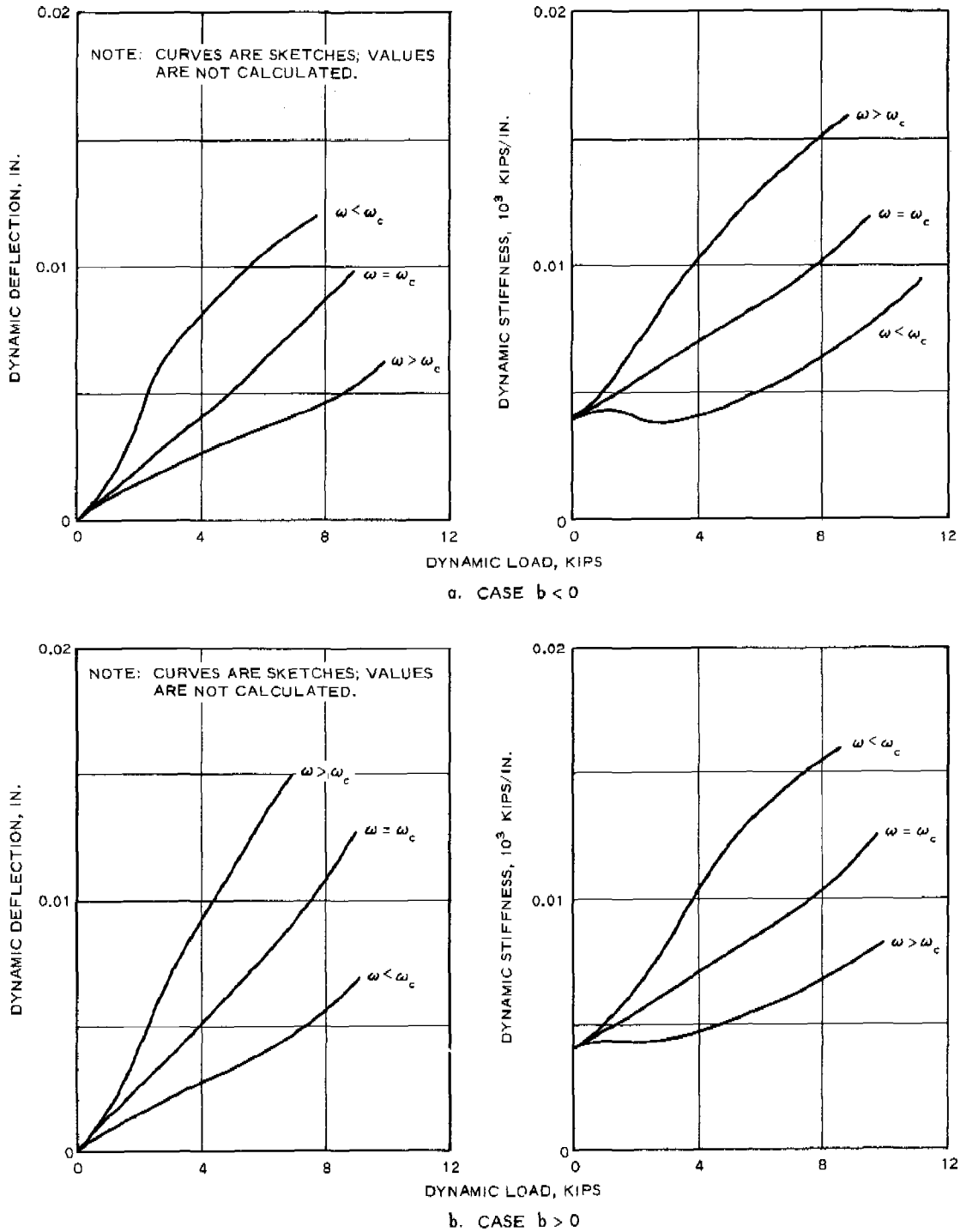
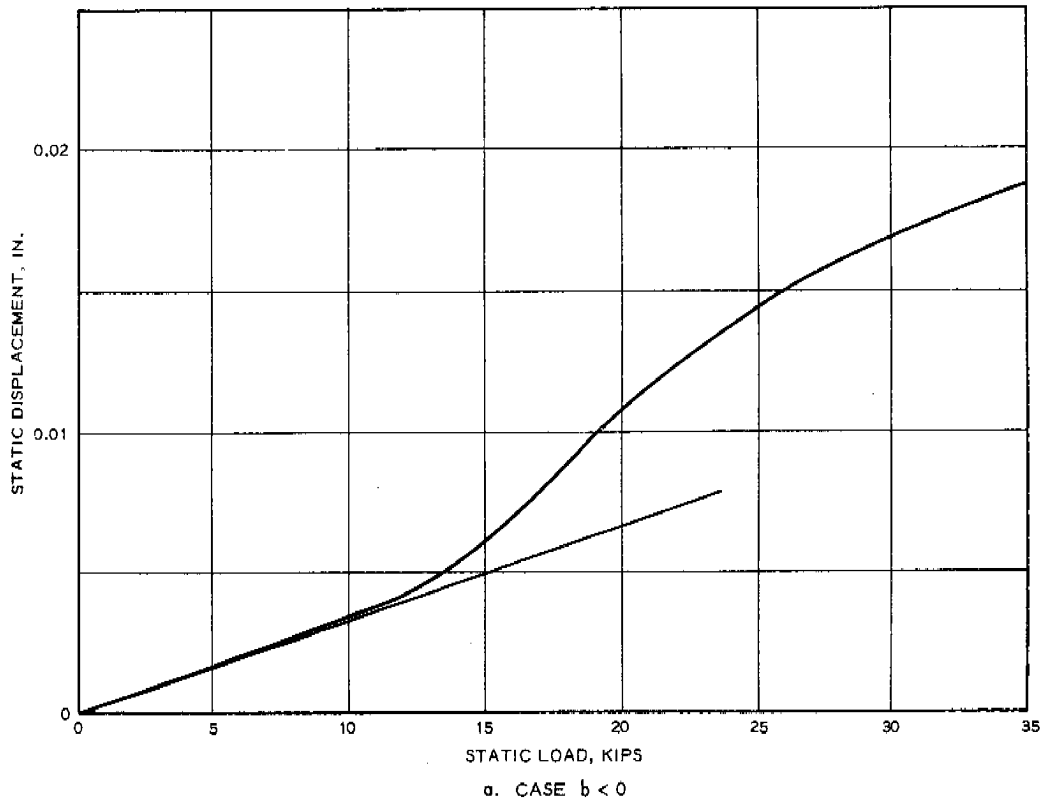
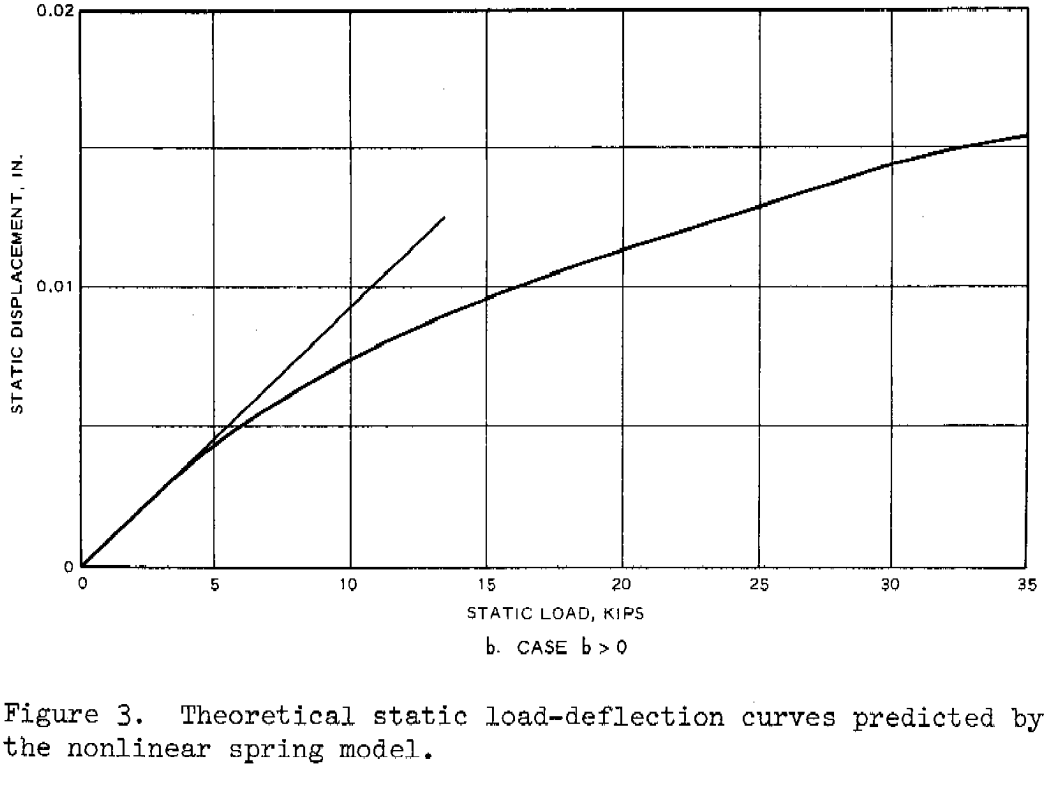


Figure 2. Theoretical dynamic load-deflection curves and dynamic stiffness curves predicted by the nonlinear spring model.



a. CASE $b < 0$



b. CASE $b > 0$

Figure 3. Theoretical static load-deflection curves predicted by the nonlinear spring model.

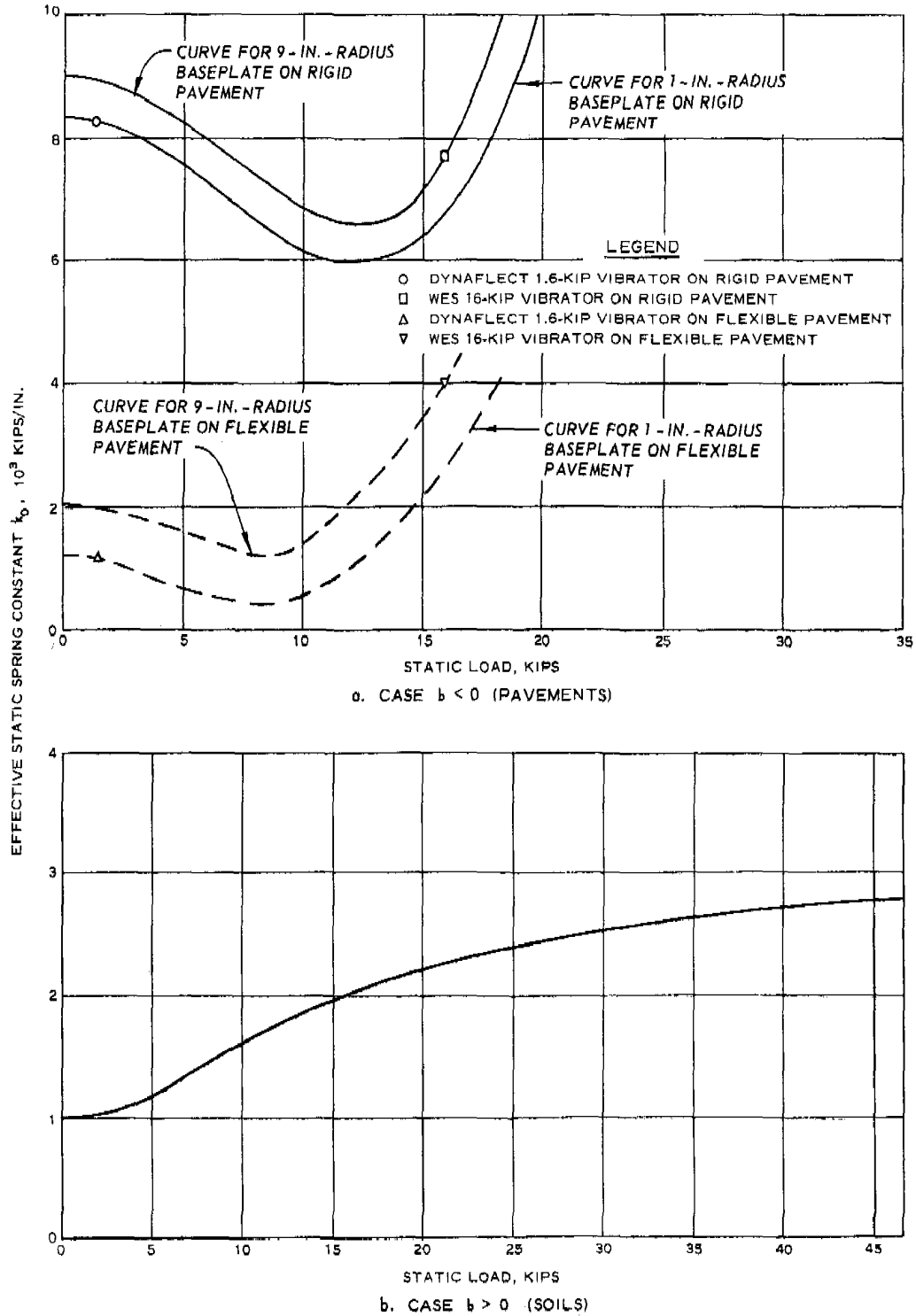


Figure 4. Theoretical dependence of the effective static spring constant on the static load.

Application of the Theory of Slender Curved Rods to the Analysis of Elastic Yarns

N. C. Huang

The Mathematics Research Center
University of Wisconsin-Madison

Abstract

A systematic formulation of the linear theory of elastic slender curved rods is presented. First, equations of equilibrium for stress resultants and bending and twisting moments are derived utilizing Serret-Frenet formula. The generalized strains are then defined according to the principle of virtual work. Constitutive equations are obtained based on the Euler-Bernoulli-Saint Venant assumptions. The theory of slender curved rods thus formulated is applied to the analysis of the extensional deformation of elastic two-ply filament yarns and continuous filament yarns, with the filaments in the yarns treated as slender curved rods. In comparison with previous works on the stress analysis of yarns, it is found that the approach adopted here has the advantage of fewer assumptions and hence could provide more accurate results and better geometrical and physical insights into the problem. In this study, the yarn elongation and filament stresses are determined for yarns with various helical angles. The effect of superposition of a twisting moment on the axial extension is also investigated.

Sponsored by the United States Army under Contract No. DA-31-124-ARO-D-462.

1. Introduction

The theory of the extension of elastic filament yarns has been investigated by a number of authors [1-4]. For purpose of simplifying the analysis, assumptions and approximations are usually imposed. Wilson and Treloar, in their study of the two-ply filament yarns, employed the theory of the finite deformation of a straight circular cylindrical rubber rod subject to combined axial extension and torsion [1]. As the filaments in a twisted yarn are in reality not straight, corrections had to be made in Wilson and Treloar's analysis for the effects of noncircular cross section and lateral pressure between filaments. In the studies on the extension of continuous filament yarns with circular cross sections [2-4], the following assumptions are imposed. (1) The cross section of each fiber is assumed to be infinitesimal and the configuration of each fiber is considered to be perfectly helical with constant radius and the same number of turns per unit length of the yarn axis. (2) The fractional contraction in yarn diameter is assumed to be uniformly distributed across the yarn. (3) The shear forces and moments acting on all faces of the yarn element are neglected. (4) The stresses at any point in the yarn are assumed to be constant in all directions at right angles to the fiber

Sponsored by the United States Army under Contract No. DA-31-124-ARO-D-462.

axis. Assumptions (2)-(4) are employed for convenience of analysis. Their validity is, however, uncertain.

In this paper, we shall establish a more accurate theory for the extension of long elastic filament yarns. The filaments in the yarn are treated as elastic slender curved rods with the center line of the rod travelling along a circular helical path. Stress analyses of yarns are made based on the theory of slender curved rods [5]. In order to give a comprehensive presentation, a survey of the formulation of the linear theory of elastic slender curved rods is first given. The rod is treated as a one-dimensional body. Equations of equilibrium are derived utilizing the Serret-Frenet formula. The strain measures are defined by the principle of virtual work in a manner similar to that used by Sanders in his formulation of nonlinear theories of shells [6]. Constitutive equations are derived based on the Euler-Bernoulli-Saint Venant assumptions.

The linear theory of slender curved rods thus formulated has been applied to the investigation of the small extension of two-ply filament yarns [7] and continuous filament yarns [8]. A survey of these studies is presented in this paper. The yarn elongation and filament stresses are analyzed for yarns with various geometries. The effect of superposition of a twisting moment on the axial extension is also discussed. In comparing our investigation with previous works [1-4], it is found that the analysis presented here has the advantage of the relaxation of assumptions and the elimination of approximations. Also, it is found that the approach adopted here can provide better geometrical and physical insights into the yarn problems.

2. Linear Theory of Slender Curved Rods

Let us consider a slender curved rod. The rectangular Cartesian coordinates of any point on the center line of the rod are

$$x_i = x_i(s) , \qquad i = 1, 2, 3, \tag{2.1}$$

where s is the arc length measured along the center line of the rod from a fixed point. The infinitesimal arc length ds satisfies

$$ds^2 = dx_i dx_i , \tag{2.2}$$

where the repeated indices represent summation. Let us denote the unit tangential, principal normal and binormal vectors by λ_i, μ_i and ν_i respectively and the principal normal curvature and the torsion of the center line by κ and τ respectively. Hence

$$\lambda_i = x_i' , \tag{2.3}$$

where prime represents the differentiation with respect to s. By Serret-Frenet formula, we have

$$\lambda_i' = \kappa\mu_i , \tag{2.4}$$

$$\mu_i' = \tau\nu_i - \kappa\lambda_i , \tag{2.5}$$

$$\nu_i' = -\tau\mu_i . \tag{2.6}$$

In the following, we shall consider only the small deformation of the rod. Therefore we can use the undeformed rod as our state of reference.

Let us denote the components in the λ_i, μ_i and ν_i directions by the subscripts λ, μ and ν respectively. The distributed force p_i and the distributed moment m_i per unit length of the rod can be

written as

$$p_i = p_\lambda \lambda_i + p_\mu \mu_i + p_\nu \nu_i \,, \tag{2.7}$$

$$m_i = m_\lambda \lambda_i + m_\mu \mu_i + m_\nu \nu_i \,. \tag{2.8}$$

The stress resultant F_i and the moment M_i acting at any cross section of the rod can be expressed as

$$F_i = F_\lambda \lambda_i + F_\mu \mu_i + F_\nu \nu_i \,, \tag{2.9}$$

$$M_i = M_\lambda \lambda_i + M_\mu \mu_i + M_\nu \nu_i \,. \tag{2.10}$$

Also, the displacement vector U_i at any point on the center line of the rod can be written as

$$U_i = u_\lambda \lambda_i + u_\mu \mu_i + u_\nu \nu_i \,. \tag{2.11}$$

Let us consider an element of the rod of length ds. The conditions of equilibrium of all forces and moments acting on the element lead to the following equations:

$$F_i' + p_i = 0 \,, \tag{2.12}$$

$$M_i' + e_{ijk} \lambda_j F_k + m_i = 0 \,, \tag{2.13}$$

where e_{ijk} is the permutation symbol. After substitutions of equations (2.7)-(2.11) into equations (2.12) and (2.13) and utilization of equations (2.4)-(2.6), we obtain the following equations of equilibrium of forces

$$F_\lambda' - \kappa F_\mu + p_\lambda = 0 \,, \tag{2.14}$$

$$F_\mu' + \kappa F_\lambda - \tau F_\nu + p_\mu = 0 \,, \tag{2.15}$$

$$F_\nu' + \tau F_\mu + p_\nu = 0 \,, \tag{2.16}$$

and the following equations of equilibrium of moments

$$M_\lambda' - \kappa M_\mu + m_\lambda = 0 , \tag{2.17}$$

$$M_\mu' + \kappa M_\lambda - \tau M_\nu - F_\nu + m_\mu = 0 , \tag{2.18}$$

$$M_\nu' + \tau M_\mu + F_\mu + m_\nu = 0 . \tag{2.19}$$

Let δu_λ, δu_μ, δu_ν be the components of virtual displacement at any point on the center line of the rod and $\delta\varphi_\lambda$, $\delta\varphi_\mu$ and $\delta\varphi_\nu$ the components of virtual rotation of any cross section. By equations of equilibrium (2.14)-(2.19), we have

$$\begin{aligned}\int_0^L [\,&\delta u_\lambda (F_\lambda' - \kappa F_\mu + p_\lambda) + \delta u_\mu (F_\mu' + \kappa F_\lambda - \tau F_\nu + p_\mu) + \delta u_\nu (F_\nu' + \tau F_\mu + p_\nu) \\ &+ \delta\varphi_\lambda (M_\lambda' - \kappa M_\mu + m_\lambda) + \delta\varphi_\mu (M_\mu' + \kappa M_\lambda - \tau M_\nu - F_\nu + m_\mu) \\ &+ \delta\varphi_\nu (M_\nu' + \tau M_\mu + F_\mu + m_\nu)]\,ds = 0 ,\end{aligned} \tag{2.20}$$

where L is the total length of the rod. After integration by parts, equations (2.20) can be written as

$$\begin{aligned}&\int_0^L (p_\lambda \delta u_\lambda + p_\mu \delta u_\mu + p_\nu \delta u_\nu + m_\lambda \delta\varphi_\lambda + m_\mu \delta\varphi_\mu + m_\nu \delta\varphi_\nu)\,ds \\ &\quad + (F_\lambda \delta u_\lambda + F_\mu \delta u_\mu + F_\nu \delta u_\nu + M_\lambda \delta\varphi_\lambda + M_\mu \delta\varphi_\mu + M_\nu \delta\varphi_\nu)\Big|_{s=0}^{s=L} \\ &\quad = \int_0^L [F_\lambda \delta(u_\lambda' - \kappa u_\mu) + F_\mu \delta(u_\mu' + \kappa u_\lambda - \tau u_\nu - \varphi_\nu) + F_\nu \delta(u_\nu' + \tau u_\mu + \varphi_\mu) \\ &\quad + M_\lambda \delta(\varphi_\lambda' - \kappa\varphi_\mu) + M_\mu \delta(\varphi_\mu' - \tau\varphi_\nu + \kappa\varphi_\lambda) + M_\nu \delta(\varphi_\nu' + \tau\varphi_\mu)]\,ds .\end{aligned} \tag{2.21}$$

The left-hand side of equation (2.21) can be identified as the external virtual work. Hence, by the principle of virtual work, the right-hand side of equation (2.21) is the internal virtual work. From equation (2.21), we may define the following strain measures for the slender curved rod:

$$\varepsilon_\lambda = u'_\lambda - \kappa u_\mu , \tag{2.22}$$

$$\gamma_\mu = u'_\mu + \kappa u_\lambda - \tau u_\nu - \varphi_\nu , \tag{2.23}$$

$$\gamma_\nu = u'_\nu + \tau u_\mu + \varphi_\mu , \tag{2.24}$$

$$\bar{\theta} = \varphi'_\lambda - \kappa \varphi_\mu , \tag{2.25}$$

$$K_\mu = \varphi'_\mu - \tau \varphi_\nu + \kappa \varphi_\lambda , \tag{2.26}$$

$$K_\nu = \varphi'_\nu + \tau \varphi_\mu . \tag{2.27}$$

We shall call ε_λ the axial strain, γ_μ and γ_ν the components of transverse shearing strain, $\bar{\theta}$ the twisting strain and K_μ and K_ν the components of bending strain.

Based on Euler-Bernoulli-Saint Venant assumptions, the displacement components at any point in the rod can be derived by the superposition of deformations due to the displacement components at the center line of the rod, the components of rotation of the cross section and the warping of the cross section introduced by the twist of the rod. Let us consider any cross section of the rod. Set the origin at the center of the cross section and two coordinate axes in the μ_i and ν_i directions. The coordinates of any point in the cross section are denoted

by μ and ν. Let $\hat{i}_s$, $\hat{i}_\mu$ and $\hat{i}_\nu$ be unit vectors in the tangential, principal normal and binormal directions respectively. The displacement vector at any point in the rod is given as

$$\hat{u} = (u_\lambda - \mu\varphi_\nu - \nu\varphi_\mu - \bar{\theta}\psi)\hat{i}_s + (u_\mu - \nu\varphi_\lambda)\hat{i}_\mu + (u_\nu + \mu\varphi_\lambda)\hat{i}_\nu , \qquad (2.28)$$

where $\psi = \psi(s, \mu, \nu)$ is a warping function associated with the twist of the rod. It can be shown that when the dimension of the cross section of the rod is small in comparison with both $1/\kappa$ and $1/\tau$, the non-vanishing components of strain at any point in the rod can be expressed as [5]

$$\varepsilon_{ss} = \varepsilon_\lambda + \nu K_\mu - \mu K_\nu + \psi\bar{\theta}' + [\tau(\nu\frac{\partial\psi}{\partial\mu} - \mu\frac{\partial\psi}{\partial\nu}) + \psi']\bar{\theta} , \qquad (2.29)$$

$$\varepsilon_{s\mu} = \frac{1}{2}[\gamma_\mu - \nu\bar{\theta} + (\frac{\partial\psi}{\partial\mu} + \kappa\psi)\bar{\theta}] , \qquad (2.30)$$

$$\varepsilon_{s\nu} = \frac{1}{2}(\gamma_\nu + \mu\bar{\theta} + \frac{\partial\psi}{\partial\nu}\bar{\theta}) . \qquad (2.31)$$

The material of the rod is considered to be linearly elastic. The stress components σ_{ss}, $\sigma_{s\mu}$ and $\sigma_{s\nu}$ are related to the strain components ε_{ss}, $\varepsilon_{s\mu}$ and $\varepsilon_{s\nu}$ by

$$\sigma_{ss} = E\varepsilon_{ss} , \qquad (2.32)$$

$$\sigma_{s\mu} = 2G\varepsilon_{s\mu} , \qquad (2.33)$$

$$\sigma_{s\nu} = 2G\varepsilon_{s\nu} , \qquad (2.34)$$

where E and G are Young's modulus and shear modulus respectively. The cross section of the rod is assumed to be doubly symmetrical with

respect to the coordinate axes. Hence, the warping function ψ is an odd function of μ and ν. The constitutive equations can be obtained by the following area integrals extended over the cross section of the rod:

$$F_\lambda = \int_A \sigma_{ss} dA = EA\varepsilon_\lambda + ED\tau\bar{\theta} , \tag{2.35}$$

$$F_\mu = \int_A \sigma_{s\mu} dA = GA\gamma_\mu , \tag{2.36}$$

$$F_\nu = \int_A \sigma_{s\nu} dA = GA\gamma_\nu , \tag{2.37}$$

$$M_\lambda = \int_A (\mu\sigma_{s\nu} - \nu\sigma_{s\mu}) dA = GJ\bar{\theta} , \tag{2.38}$$

$$M_\mu = \int_A \nu\sigma_{ss} dA = EI_\mu K_\mu , \tag{2.39}$$

$$M_\nu = \int_A \mu\sigma_{ss} dA = EI_\nu K_\nu , \tag{2.40}$$

where A is the cross-sectional area, $I_\mu = \int_A \nu^2 dA$, $I_\nu = \int_A \mu^2 dA$ are moments of inertia of the cross section about the μ and ν axes respectively,

$$J = \int_A (\mu \frac{\partial\psi}{\partial\nu} - \nu \frac{\partial\psi}{\partial\mu} + \mu^2 + \nu^2) dA \tag{2.41}$$

is the torsional constant of the cross section and $D = I_\mu + I_\nu - J$.

Equations (2.14)-(2.19), (2.22)-(2.27) and (2.35)-(2.40) are the governing equations of the small deformation of slender curved rods. The boundary conditions can be obtained from equation (2.21). We find that at

both ends of the rod, it is necessary to prescribe (1) F_λ or u_λ, (2) F_μ or u_μ, (3) F_ν or u_ν, (4) M_λ or φ_λ, (5) M_μ or φ_μ, (6) M_ν or φ_ν. We shall employ these field equations to analyze the deformation of elastic yarns subject to axial extension in the following sections.

3. Extension of Elastic Two-Ply Filament Yarns

Let us consider a two-ply filament yarn subject to an axial force $\bar{F}$ and a twisting moment $\bar{M}$. We consider $\bar{M}$ to be positive if it is in the direction of the twist of the yarn. The cross section of each filament in the undeformed state is assumed to be circular with radius a. The center line of the filament is regarded as helical with the length of one turn of the twist measured along the axis of the yarn as $h = 2\pi k$ where k is a constant. Let us denote the distance measured from the yarn axis to any point on the center line of the filament by r and the helical angle of the center line by θ. The principal normal curvature and the torsion of the center line of the filament are found to be

$$\kappa = \frac{1}{\rho}\sin\theta, \qquad \tau = \frac{1}{\rho}\cos\theta, \tag{3.1}$$

where $\rho = (r^2 + k^2)^{\frac{1}{2}}$ and $\theta = \tan^{-1}\frac{r}{k}$. The unit vector $\hat{\lambda}$ makes an angle θ with the yarn axis. The unit vector $\hat{\mu}$ is in the radial direction and toward the yarn axis. Hence the angle between $\hat{\mu}$ and the yarn axis is $\pi/2$.

The filament is considered to be a one-dimensional body of infinite length. We shall apply the field equations of slender curved rods to analyze the deformation of the filament. In the equations of equilibrium of our problem, the derivatives of the stress resultant and moment with respect to the arc length of the filament must vanish. Hence, by equations (2.14)-(2.19), it is found that

$$F_\mu \kappa - p_\lambda = 0 , \tag{3.2}$$

$$F_\lambda \kappa - F_\nu \tau + p_\mu = 0 , \tag{3.3}$$

$$F_\mu \tau + p_\nu = 0 , \tag{3.4}$$

$$M_\mu \kappa - m_\lambda = 0 , \tag{3.5}$$

$$M_\lambda \kappa - M_\nu \tau - F_\nu + m_\mu = 0 , \tag{3.6}$$

$$M_\mu \tau + F_\mu + m_\nu = 0 . \tag{3.7}$$

The strain-displacement relations are given by equations (2.22)-(2.27) and the constitutive relations are given by equations (2.35)-(2.40) with $A = \pi a^2$, $I_\mu = I_\nu = I = \frac{\pi}{4} a^4$, $J = \frac{\pi}{2} a^4$ and $D = 0$.

The overall equilibrium of the internal forces and the applied force $\bar{F}$ requires that

$$\bar{F} = 2(F_\lambda \cos\theta + F_\nu \sin\theta) . \tag{3.8}$$

Similarly, the overall equilibrium in moments requires that

$$\bar{M} = 2(M_\lambda \cos\theta + M_\nu \sin\theta + F_\lambda r \sin\theta - F_\nu r \cos\theta) . \tag{3.9}$$

Since $\hat{\mu}$ is in the radial direction and the yarn is long, the components u_μ and φ_μ must be independent of the arc length of the

filament. If we set the origin $s = 0$ at the mid-point of the entire yarn, we may write

$$u_\lambda = c_1 s, \quad u_\nu = c_2 s, \quad \varphi_\lambda = c_3 s, \quad \varphi_\nu = c_4 s , \tag{3.10}$$

where c_1, c_2, c_3 and c_4 are constants to be determined. After substitutions, we obtain

$$F_\lambda = EA(c_1 - u_\mu \kappa) , \tag{3.11}$$

$$F_\mu = GA(c_1 \kappa - c_2 \tau - c_4) s , \tag{3.12}$$

$$F_\nu = GA(c_2 + u_\mu \tau + \varphi_\mu) , \tag{3.13}$$

$$M_\lambda = GJ(c_3 - \varphi_\mu \kappa) , \tag{3.14}$$

$$M_\mu = EI(c_3 \kappa - c_4 \tau) s , \tag{3.15}$$

$$M_\nu = EI(c_4 + \varphi_\mu \tau) . \tag{3.16}$$

Since the yarn is assumed to be long, p_λ and m_λ are constant. From equations (3.2) and (3.5), it is seen that F_μ and M_μ are also constant. However, equations (3.12) and (3.15) indicate that F_μ and M_μ are proportional to s. Therefore, we conclude that $F_\mu = M_\mu = 0$. From equations (3.12) and (3.15), we find that

$$c_3 = \frac{\tau}{\kappa}(c_1 \kappa - c_2 \tau) , \tag{3.17}$$

$$c_4 = c_1 \kappa - c_2 \tau , \tag{3.18}$$

and from equations (3.2), (3.4), (3.5) and (3.7), we find that $p_\lambda = p_\nu = m_\lambda = m_\nu = 0$. Hence p_μ is the only nonvanishing contact pressure between the filaments, i.e., the common normal to the line of

contact of the two filaments is in the radial direction. We conclude that the line of contact must coincide with the yarn axis and the two filaments wind about the yarn axis following helical paths as shown in Figure 1. Accordingly, we have

$$r = a, \quad \rho = a \csc \theta . \tag{3.19}$$

In the deformed state, the filaments contact by a curved surface. The analysis of this type of contact problem is difficult. In our analysis, we assume that the contribution to u_μ due to the elastic contact of filaments is proportional to p_μ. Hence, we have

$$u_\mu = \frac{\sigma a}{AE} \Gamma_\lambda - \zeta p_\mu , \tag{3.20}$$

where σ is Poisson's ratio and ζ is a positive constant. The first term in equation (3.20) is due to the contraction in diameter of the filament through Poisson's effect. The distributed moment m_μ can be interpreted as the moment caused by friction due to the relative motion of the filaments. The magnitude of m_μ depends on the width of the contact surface. Under the assumption of small extension, we may set $m_\mu = 0$.

After substitutions, equations (3.3)-(3.6) can be written as

$$(\kappa + \frac{1}{\zeta} \frac{\sigma a}{AE}) \Gamma_\lambda - GA\tau\gamma_\nu - \frac{1}{\zeta} u_\mu = 0 , \tag{3.21}$$

$$M_\lambda \kappa - M_\nu \tau - GA\gamma_\nu = 0 . \tag{3.22}$$

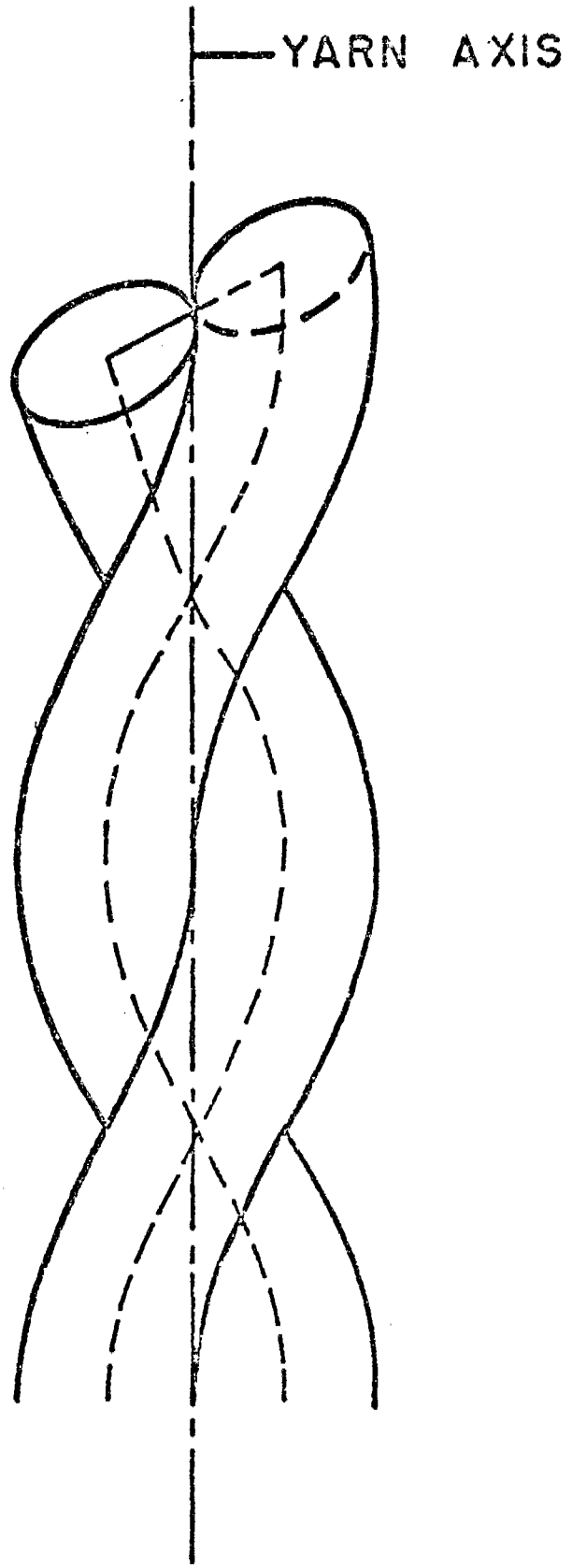


Figure 1. Configuration of a Two Ply Filament Yarn

Let u_θ and u_z be the displacement components of the filament in the tangential and axial directions in a cylindrical polar coordinates system. It is found that

$$u_\theta = \frac{1}{\rho}(ru_\lambda - ku_\nu) = (c_1\kappa - c_2\tau)s , \tag{3.23}$$

$$u_z = \frac{1}{\rho}(ku_\lambda + ru_\nu) = (c_1\tau + c_2\kappa)s . \tag{3.24}$$

Hence the rotation of the cross section of the yarn about its axis is

$$\omega_z = \frac{\rho}{r}(c_1\kappa - c_2\tau)s . \tag{3.25}$$

Let us denote the extension and torsion of the yarn per unit length by u^* and ω^* respectively. We have

$$u^* = c_1 + c_2 \tan\theta , \tag{3.26}$$

$$\omega^* = \frac{1}{r}(c_1 \tan\theta - c_2) . \tag{3.27}$$

By equations (3.17), (3.18), (3.26) and (3.27), we can express c_1, c_2, c_3 and c_4 in terms of u^* and ω^*. By equations (3.11), (3.13), (3.14), (3.16), (2.37) and (2.24), we are able to express F_λ, F_ν, M_λ and M_ν as linear functions of u^*, ω^*, u_μ and γ_ν with coefficients dependent on θ. A set of simultaneous linear equations of u^*, ω^*, u_μ and γ_ν can then be derived from equations (3.21), (3.22), (3.8) and (3.9). Let us introduce the following dimensionless quantities:

$$\alpha = k/a,\ \xi = E\zeta,\ u = 2\pi aEu_\mu/\bar{F},\ U^* = 2\pi a^2Eu^*/\bar{F},\ V^* = 2\pi a^3E\omega^*/\bar{F} ,$$

$$\gamma^* = 2\pi a^2E\gamma_\nu/\bar{F},\ m = \bar{M}/(\bar{F}a),\ f = 2F_\lambda/\bar{F} . \tag{3.28}$$

The helical angle θ is related to the pitch parameter α by $\theta = \cot^{-1}\alpha$. Our governing equations can be written in a form of $\underset{\sim}{A}\underset{\sim}{Y} = \underset{\sim}{Z}$ where $\underset{\sim}{Y} = [U^* V^* u \gamma^*]^T$, $\underset{\sim}{Z} = [00lm]^T$ and $\underset{\sim}{A}$ is a four by four matrix whose elements depend on θ.

For any given values of σ, α, ξ and m, we can find U^*, V^*, u and γ^* by solving simultaneous equations. The axial force carried by each filament can be found from equation (3.11) as

$$f = U^* \cos^2\theta + V^* \sin\theta \cos\theta - u \cos^2\theta . \qquad (3.29)$$

In our study, we adopt $\sigma = 0.4$ and $\xi = 1.07$. Two problems are being considered here. In the first problem, the ends of the yarn are clamped. This corresponds to the case of tension test of yarns. In this case, $V^* = 0$ and U^*, u, γ^* and m are unknown. The calculated values of U^*, m and f are plotted as functions of α in Figure 2 by the solid lines. Note that when α approaches infinity, both U^* and f approach a limit one. When α is large, the filament are nearly straight. When α is small, the elongation of the yarn is essentially governed by the change of the helical angle. Between these two limiting cases, the value of m may reach a maximum value as shown in Figure 2.

In the second problem, the ends of the yarn are free to rotate. This case occurs when a vertical yarn is fixed at the upper end and extended by a weight attached to the lower end. In this problem, $m = 0$ and

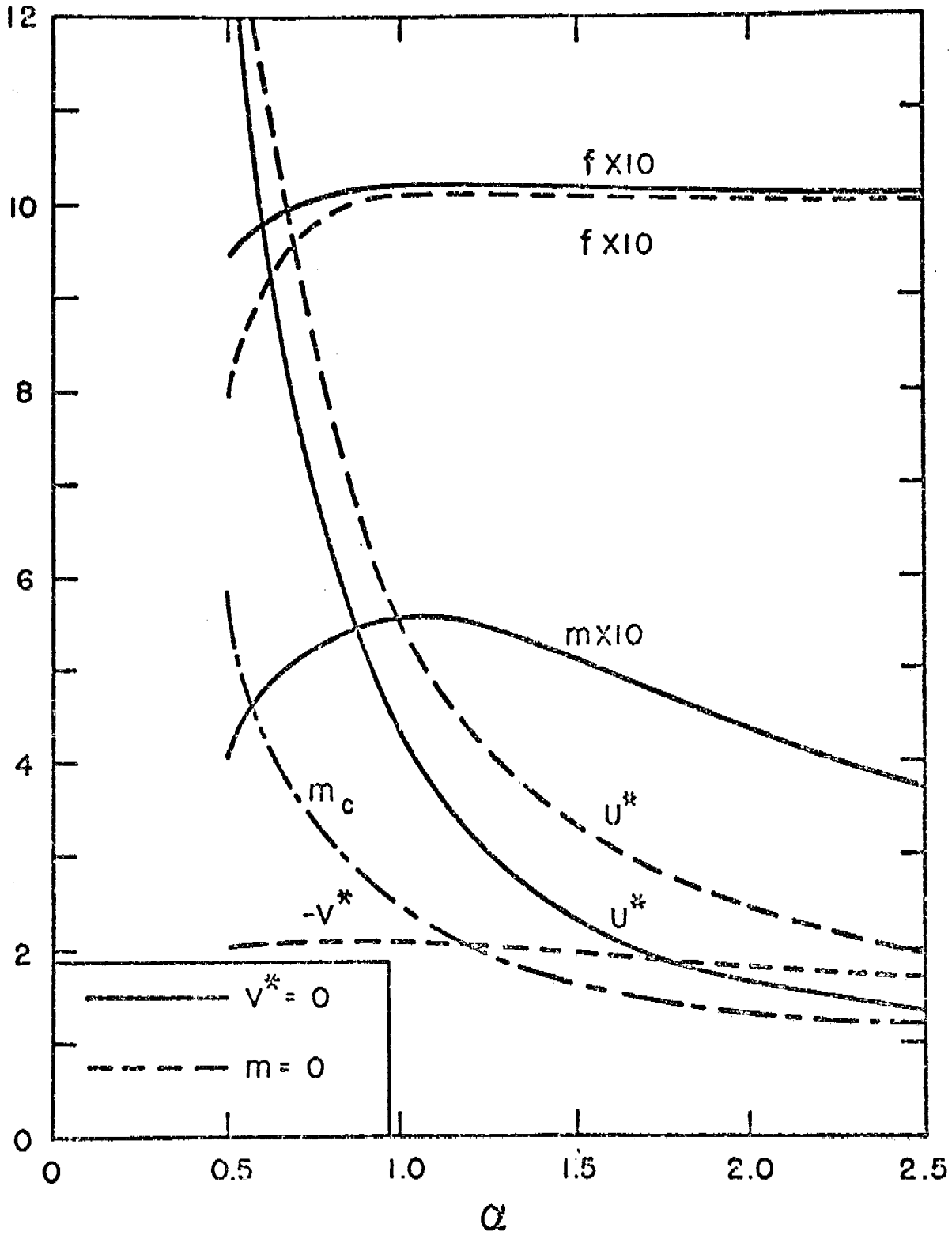


Figure 2. U^*, V^*, m, f and m_c Versus α Curves for $\sigma = 0.4$ and $\xi = 1.07$.

U^*, V^*, u and γ are unknown. The calculated values of U^*, V^* and f are plotted against α in Figure 2 by the dotted lines. The negative value of V^* shows that an untwist occurs during extension. When α approaches infinity, both U^* and f approach a limit one and V^* approaches zero. For given values of σ, α and ξ, the quantities U^*, V^*, u and γ^* are linear functions of m. It is found that U^* decreases with increasing m. When m is sufficiently large, U^* becomes negative. The value of m for $U^* = 0$ is denoted by m_c. It is found that m_c decreases with increasing α as shown in Figure 2.

4. Extension of Elastic Continuous Filament Yarns

Let us consider an idealized infinitely long yarn of circular cross section with radius R subject to an axial force $\bar{F}$ and a twisting moment $\bar{M}$ at both ends as shown in Figure 3a. The yarn is composed of a large number of helical fibers. The length of one turn of twist of all fibers is regarded as constant and is denoted by $h = 2\pi k$. We shall treat each fiber element as a slender curved rod of helical configuration. The cross-sectional area of the yarn element cut by a plane perpendicular to the yarn axis is $rdrd\varphi$. Hence the cross-sectional area of the element perpendicular to its own axis is $dA = r \cos \theta drd\varphi$, where θ is the helical angle of the element. The principal normal curvature and the torsion of the element are again given by equation (3.1). The components

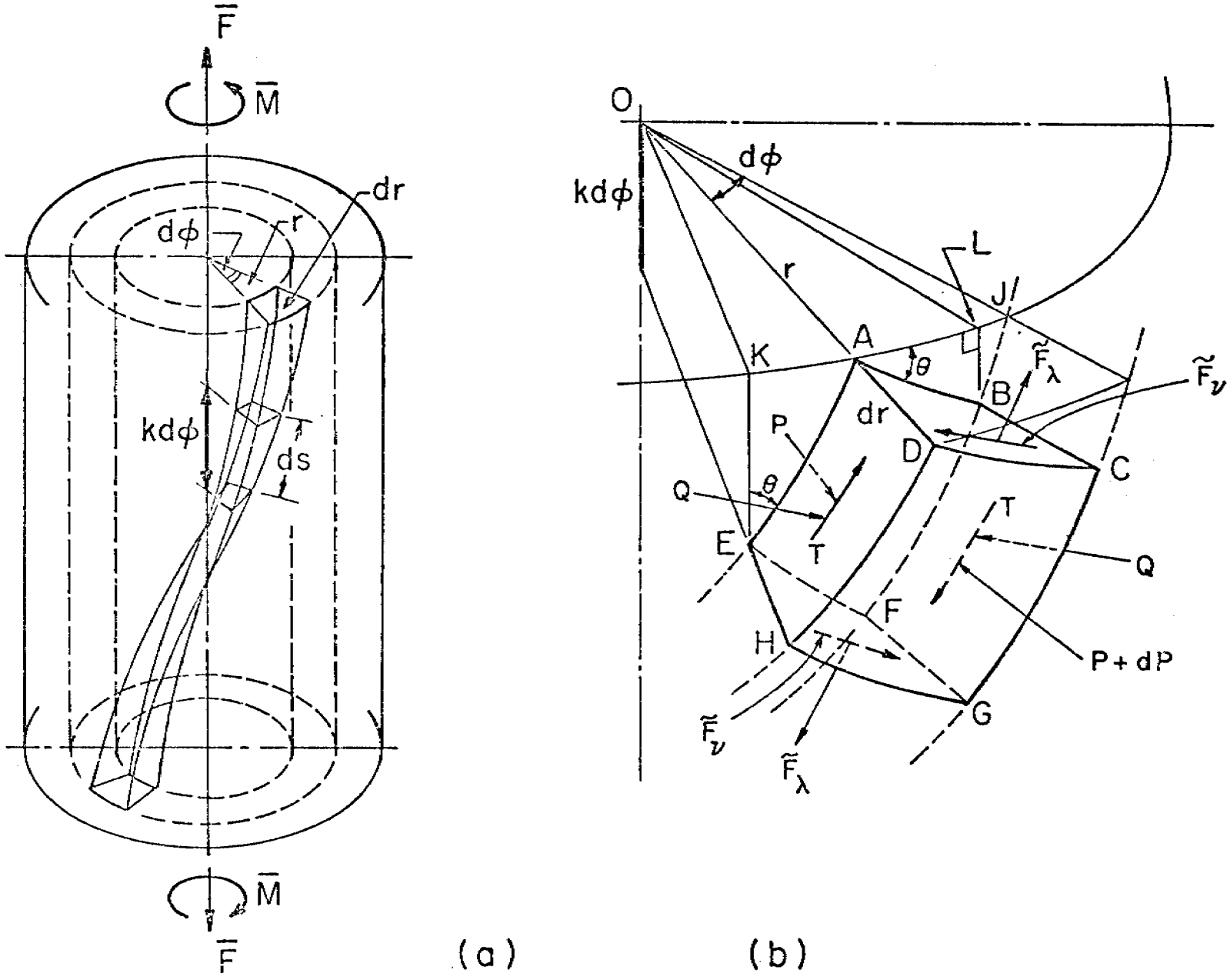


Fig. 3. The Yarn Geometry and Stresses Acting on an Element.

of the stress resultant acting on the cross section of the yarn element are denoted by dF_λ, dF_μ and dF_ν and the components of the moment acting on the cross section are denoted by dM_λ, dM_μ and dM_ν. The components of the distributed force and the distributed moment per unit length of the element are dp_λ, dp_μ, dp_ν, dm_λ, dm_μ and dm_ν. Put $\tilde{(\,)} = \frac{d}{dA}(\,)$. Equations of equilibrium for a long yarn are similar to equations (3.2)-(3.7). They are

$$\tilde{F}_\mu \kappa - \tilde{p}_\lambda = 0 , \tag{4.1}$$

$$\tilde{F}_\lambda \kappa - \tilde{F}_\nu \tau + \tilde{p}_\mu = 0 , \tag{4.2}$$

$$\tilde{F}_\mu \tau + \tilde{p}_\nu = 0 , \tag{4.3}$$

$$\tilde{M}_\mu \kappa - \tilde{m}_\lambda = 0 , \tag{4.4}$$

$$\tilde{M}_\lambda \kappa - \tilde{M}_\nu \tau - \tilde{F}_\nu + \tilde{m}_\mu = 0 , \tag{4.5}$$

$$\tilde{M}_\mu \tau + \tilde{F}_\mu + \tilde{m}_\nu = 0 . \tag{4.6}$$

The strain-displacement relations are again given by equations (2.22)-(2.27). Based on the same reason as that used in the analysis of the two-ply filament yarns, it is found that equation (3.10) is also valid in the case of continuous filament yarns.

Let us denote the normal compressive stresses in the μ and ν directions by P and Q, the normal strain components in the μ and ν directions by ε_μ and ε_ν and the shearing stress and shearing strain corresponding to the λ, ν directions by T and $\varepsilon_{\lambda\nu}$. Here, we assume

that there is no slipping between fibers. Hence T is caused by friction. The stresses acting on an infinitesimal yarn element are shown in Figure 3b. As we shall see later, all stress components not shown in Figure 3b are actually zero. From Figure 3b, it is found that $\tilde{p}_\mu$ and $\tilde{m}_\mu$ are related to P, Q and T by

$$\tilde{p}_\mu = \frac{dP}{dr} + \frac{1}{r}(P - Q\cos^2\theta)\,, \tag{4.7}$$

$$\tilde{m}_\mu = T\,. \tag{4.8}$$

The cross section of the yarn element is infinitesimal. It can be regarded as doubly symmetrical. The fiber is considered to be elastic and transversely isotropic. We have the following constitutive relations:

$$\tilde{F}_\lambda = a_1\varepsilon_\lambda + a_2(\varepsilon_\mu + \varepsilon_\nu) + Er_d^2\tau\bar{\theta}\,, \tag{4.9}$$

$$P = -(a_2\varepsilon_\lambda + a_3\varepsilon_\mu + a_4\varepsilon_\nu)\,, \tag{4.10}$$

$$Q = -(a_2\varepsilon_\lambda + a_4\varepsilon_\mu + a_3\varepsilon_\nu)\,, \tag{4.11}$$

$$\tilde{F}_\mu = \frac{1}{2}a_5\gamma_\mu\,, \tag{4.12}$$

$$\tilde{F}_\nu = \frac{1}{2}a_5\gamma_\nu\,, \tag{4.13}$$

$$T = a_5\varepsilon_{\lambda\nu}\,, \tag{4.14}$$

where $a_i (i = 1, 2, \ldots, 5)$ are material constants, $E = a_1 - 2a_2^2/(a_3 + a_4)$ is the elastic modulus of the fiber in the tangential direction, $r_d^2 = (I_\mu + I_\nu - J)/A$, I_μ and I_ν are the moments of inertia of the cross section and J is the torsional constant. In comparing equation (4.9) with (2.35), it is found that an additional term of $a_2(\varepsilon_\mu + \varepsilon_\nu)$ is added

to include the effect of the normal strain components in the transverse direction. The constitutive relations for $\tilde{M}_\lambda$, $\tilde{M}_\mu$ and $\tilde{M}_\nu$ are

$$\tilde{M}_\lambda = \frac{1}{2} a_5 r_\lambda^2 \bar{\theta} , \tag{4.15}$$

$$\tilde{M}_\mu = E r_\mu^2 K_\mu , \tag{4.16}$$

$$\tilde{M}_\nu = E r_\nu^2 K_\nu , \tag{4.17}$$

where r_λ, r_μ and r_ν are radii of gyration corresponding to the moments of inertia J, I_μ and I_ν respectively.

By a similar procedure as what we used in the analysis of two-ply filament yarns, we can prove that $\tilde{F}_\mu = \tilde{M}_\mu = \tilde{p}_\lambda = \tilde{m}_\lambda = 0$. We can also show that equations (3.26) and (3.27) still hold in this problem. By substitutions, we obtain

$$\tilde{F}_\lambda = a_1(u^* \cos^2\theta + r\omega^* \sin\theta \cos\theta - \frac{u_\mu}{\rho} \sin\theta) + a_2(\varepsilon_\mu + \varepsilon_\nu) + E \frac{r_d^2}{\rho}(\omega^* \cos^3\theta - \frac{\varphi_\mu}{\rho} \sin\theta \cos\theta) , \tag{4.18}$$

$$P = -a_2(u^* \cos^2\theta + r\omega^* \sin\theta \cos\theta - \frac{u_\mu}{\rho} \sin\theta) - a_3\varepsilon_\mu - a_4\varepsilon_\nu , \tag{4.19}$$

$$Q = -a_2(u^* \cos^2\theta + r\omega^* \sin\theta \cos\theta - \frac{u_\mu}{\rho} \sin\theta) - a_4\varepsilon_\mu - a_3\varepsilon_\nu , \tag{4.20}$$

$$\tilde{F}_\nu = \frac{1}{2} a_5(u^* \sin\theta \cos\theta - r\omega^* \cos^2\theta + \frac{u_\mu}{\rho} \cos\theta + \varphi_\mu) , \tag{4.21}$$

$$\tilde{M}_\lambda = \frac{1}{2} a_5 r_\lambda^2(\omega^* \cos^2\theta - \frac{\varphi_\mu}{\rho} \sin\theta) , \tag{4.22}$$

$$\tilde{M}_\nu = E r_\nu^2(\omega^* \sin\theta \cos\theta + \frac{\varphi_\mu}{\rho} \cos\theta) . \tag{4.23}$$

From equations (4.2), (4.5), (4.7), (4.8), (4.18), (4.22) and (4.23), we obtain

$$\frac{1}{\rho}\sin\theta\left[a_1\left(u^* \cos^2\theta + r\omega^* \sin\theta\cos\theta - \frac{u_\mu}{\rho}\sin\theta\right) + a_2(\varepsilon_\mu + \varepsilon_\nu)\right]$$

$$- \frac{T}{\rho}\cos\theta + \frac{dP}{dr} + \frac{1}{r}(P - \varphi\cos^2\theta) = 0 \ . \qquad (4.24)$$

In the derivation of equation (4.24), we consider that the cross section of the yarn element is infinitesimal and neglect those terms involving r_d^2, r_λ^2 or r_ν^2.

When the ends of the yarn are clamped, under a pure extension, the helical angle of fibers decreases and separation between fibers may appear in the binormal direction in the region $r > \bar{r}$. However in the central region $r < \bar{r}$, this separation does not occur. In this region, it is possible for us to find ε_μ, ε_ν and $\varepsilon_{\lambda\nu}$ in terms of u^*, ω^* and u_μ by the transformation of the strain tensor from a rectangular Cartesian coordinates system to the (λ, μ, ν) system. It is found that in the region without separation of fibers,

$$\varepsilon_\lambda = -\frac{u}{r}\sin^2\theta + u^* \cos^2\theta + r\omega^* \sin\theta\cos\theta \ , \qquad (4.25)$$

$$\varepsilon_\mu = -\frac{du_\mu}{dr} \ , \qquad (4.26)$$

$$\varepsilon_\nu = -\frac{u_\mu}{r}\cos^2\theta + u^* \sin^2\theta - r\omega^* \sin\theta\cos\theta \ , \qquad (4.27)$$

$$\varepsilon_{\lambda\nu} = \frac{u_\mu}{r}\sin\theta\cos\theta + u^* \sin\theta\cos\theta + \frac{r}{2}\omega^*(\sin^2\theta - \cos^2\theta) \ . \qquad (4.28)$$

Denote the ratio of the cross-sectional area of the void in the yarn to the total cross-sectional area of the yarn by γ. Let us introduce the following dimensionless quantities:

$$\xi = r/R,\ \bar{\xi} = \bar{r}/R,\ c = k/R,\ \eta(\xi) = (c^2 + \xi^2)^{\frac{1}{2}},\ \eta_1 = \eta(1),\ \bar{\eta} = \eta(\bar{\xi}),\ \alpha_i = a_i/E \quad (i = 1, 2, \ldots, 5),$$

$$f = \bar{F}/[2(1 - \gamma)\pi E R^2],\ m_t = \bar{M}/(\bar{F}R),\ u = u_\mu/(fR),\ \varepsilon = \varepsilon_\lambda/f,\ f_\lambda = \tilde{F}_\lambda/(Ef),$$

$$U^* = u^*/f,\ V^* = \omega^* R/f,\ p = P/(Ef),\ q = Q/(Ef),\ t = T/(Ef),\ (\ )' = \frac{d}{d\xi}(). \quad (4.29)$$

Note that $U^* = 2E/E_y$, where E_y is the overall elastic modulus of the yarn. It is found that $\alpha_1 = 1 + 2\alpha_2^2/(\alpha_3 + \alpha_4)$.

The governing differential equations for the deformation of the yarn in the region without separation can be derived from equation (4.14), (4.19), (4.20) and (4.24)-(4.27) as

$$u' + f_{11}u + f_{12}p + g_{11}U^* + g_{12}V^* = 0, \quad (4.30)$$

$$p' + f_{21}u + f_{22}p + g_{21}U^* + g_{22}V^* = 0, \quad (4.31)$$

where f_{ij} and g_{ij} $(i = 1, 2$ and $j = 1, 2)$ are functions of α_i, c, η and ξ. After u is determined, the fiber stress f_λ can be determined by equations (4.18), (4.26) and (4.27) and the contact pressure q can be determined by equation (4.20), (4.6) and (4.27).

For $r > \bar{r}$, separation of fibers occurs in the binormal direction. In this region, equation (4.27) and (4.28) are no longer valid. However, we may set $Q = T = 0$. It is found that equations (4.18), (4.19), (4.25)

and (4.26) are still valid but equations (4.20), (4.7), (4.8) and (4.24) must be replaced by the following equations:

$$a_2(u^* \cos^2\theta + r\omega^* \sin\theta \cos\theta - \frac{u_\mu}{\rho} \sin\theta) + a_4\varepsilon_\mu + a_3\varepsilon_\nu = 0 , \quad (4.32)$$

$$\tilde{p}_\mu = \frac{dP}{dr} + \frac{P}{r} , \quad (4.33)$$

$$\tilde{m}_\mu = 0 , \quad (4.34)$$

$$\frac{1}{\rho} \sin\theta[a_1(u^* \cos^2\theta + r\omega^* \sin\theta\cos\theta - \frac{u_\mu}{\rho} \sin\theta) + a_2(\varepsilon_\mu + \varepsilon_\nu)] + \frac{dP}{dr} + \frac{P}{r} = 0. \quad (4.35)$$

The governing differential equations can be obtained from equations (4.19), (4.32), (4.26) and (4.35). These equations can be expressed in the same form as equations (4.30) and (4.31). After u is determined, f_λ can be found from equations (4.18), (4.26) and (4.32).

In order to solve for the differential equations (4.30) and (4.31), we need the boundary conditions at $r = 0$ and $r = R$ and the continuity conditions at $r = \bar{r}$. These conditions are

$$u_\mu(0) = P(R) = 0 , \quad (4.36)$$

$$u_\mu(\bar{r}^-) = u_\mu(\bar{r}^+) , \quad (4.37)$$

$$P(\bar{r}^-) = P(\bar{r}^+) . \quad (4.38)$$

These conditions can be easily expressed in terms of u and u'. The correct choice of $\bar{\xi}$ must make $q(\bar{\xi}^-) = 0$. This can be done by a cut-and-try process.

In equations (4.30) and (4.31), two unknown constants U^* and V^* are involved. These constants are determined by the overall

equilibrium conditions of the yarn. From Figure 3b, it is found that the applied axial force $\bar{F}$ and twisting moment $\bar{M}$ can be expressed by the following integrals:

$$\bar{F} = 2\pi(1-\gamma)\int_0^R [(\tilde{F}_\lambda \cos\theta + \tilde{F}_\nu \sin\theta)\cos\theta + (T\cos\theta - Q\sin\theta)\sin\theta]rdr , \tag{4.39}$$

$$\bar{M} = 2\pi(1-\gamma)\int_0^R [(\tilde{M}_\lambda \cos\theta + \tilde{M}_\nu \sin\theta + \tilde{F}_\lambda r\sin\theta - \tilde{F}_\nu r\cos\theta)\cos\theta + (Tr\sin\theta + Qr\cos\theta)\sin\theta]rdr . \tag{4.40}$$

In the derivation of equations (4.39) and (4.40), the rule of mixtures of γ-diluted single fiber property has been employed. By substitutions, equations (4.39) and (4.40) can be written as

$$\int_0^{\bar{\xi}} H_1 u d\xi + \int_{\bar{\xi}}^1 H_2 u d\xi + c_{11}U^* + c_{12}V^* + F_1 u(\bar{\xi}) + G_1 u(1) = 1 , \tag{4.41}$$

$$\int_0^{\bar{\xi}} H_3 u d\xi + \int_{\bar{\xi}}^1 H_4 u d\xi + c_{21}U^* + c_{22}V^* + F_2 u(\bar{\xi}) + G_2 u(1) = m_t , \tag{4.42}$$

where $H_i (i = 1, 2, 3, 4)$ are functions of α_i, ξ and η, c_{ij} $(i = 1, 2$ and $j = 1, 2)$, F_1 and F_2 are constants dependent on α_i, c, $\bar{\xi}$, $\bar{\eta}$ and η_1 and G_1 and G_2 are constants dependent on α_i, c and η_1.

Numerical solution is first sought for the case of yarns with clamped ends. In this case, $V^* = 0$. In our calculation, we use $\alpha_1 = 1.16$, $\alpha_2 = 0.4$, $\alpha_3 = 1.2$, $\alpha_4 = 0.8$ and $\alpha_5 = 0.3$.

In Figure 4, the nondimensional radial displacement and the fiber stress are plotted against the radial coordinate for different values of c as indicated on the curves. The positions $\xi = \bar{\xi}$ are shown by dark circles. Note that in the region without separation of fibers, $u(\xi)$ is approximately proportional to ξ. However, this proportionality does not appear in the region with separation of fibers. As we may expect, the maximum fiber stress occurs at the center of the yarn. It is found that when c increases, the helical angle of the yarn decreases and hence both the radial displacement and the maximum fiber stress decreases. The contact pressure $p(\xi)$ and $q(\xi)$ and the shearing stress $t(\xi)$ are plotted against ξ in Figure 5. At the center of the yarn, $p(0) = q(0)$ and $t(0) = 0$. In the region of the separation of fibers, $q(\xi) = t(\xi) = 0$.

The normalized axial strain U^* and the twisting moment at the fixed end m_t are plotted against c in Figure 6. When c approaches zero, the elongation of the yarn is dominated by the change of the helical angle and hence both U^* and m_t approach infinity. On the other hand, when c approaches infinity, all fibers become straight. Hence U^* approaches 2 and m_t approaches zero.

The relative modulus (RM) is defined as the ratio E_y/E. In Figure 7, the normalized relative modulus with respect to the modulus of a yarn of helical angle $10°$ is plotted against the helical angle of the yarn. The curve based on a $\cos^2\alpha$ approximate theory [2] and the curve

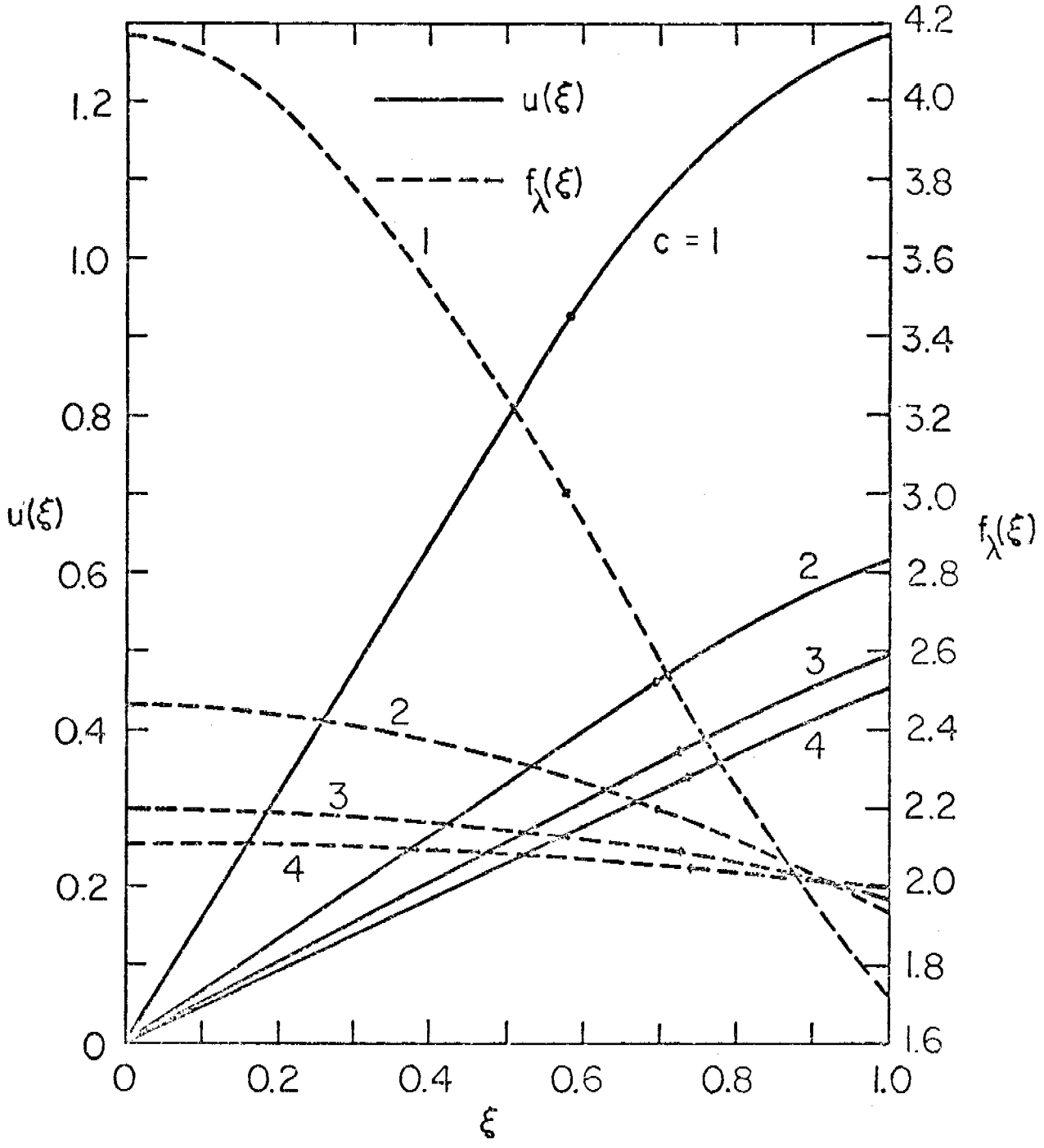


Fig. 4. $u(\xi)$ and $f_\lambda(\xi)$ Curves for Various Values of c.

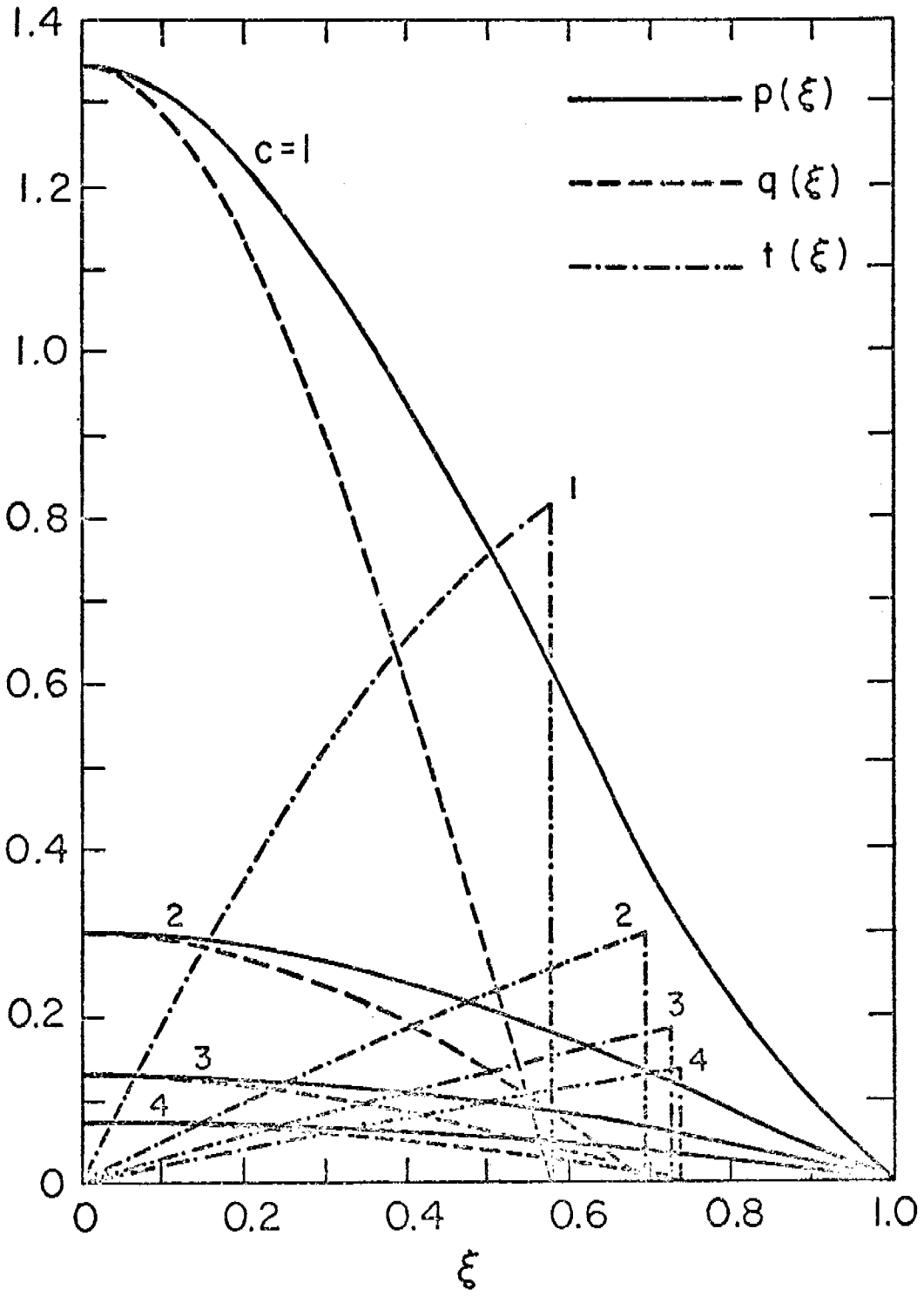


Fig. 5. $p(\xi)$, $q(\xi)$ and $t(\xi)$ Curves for Various Values of c.

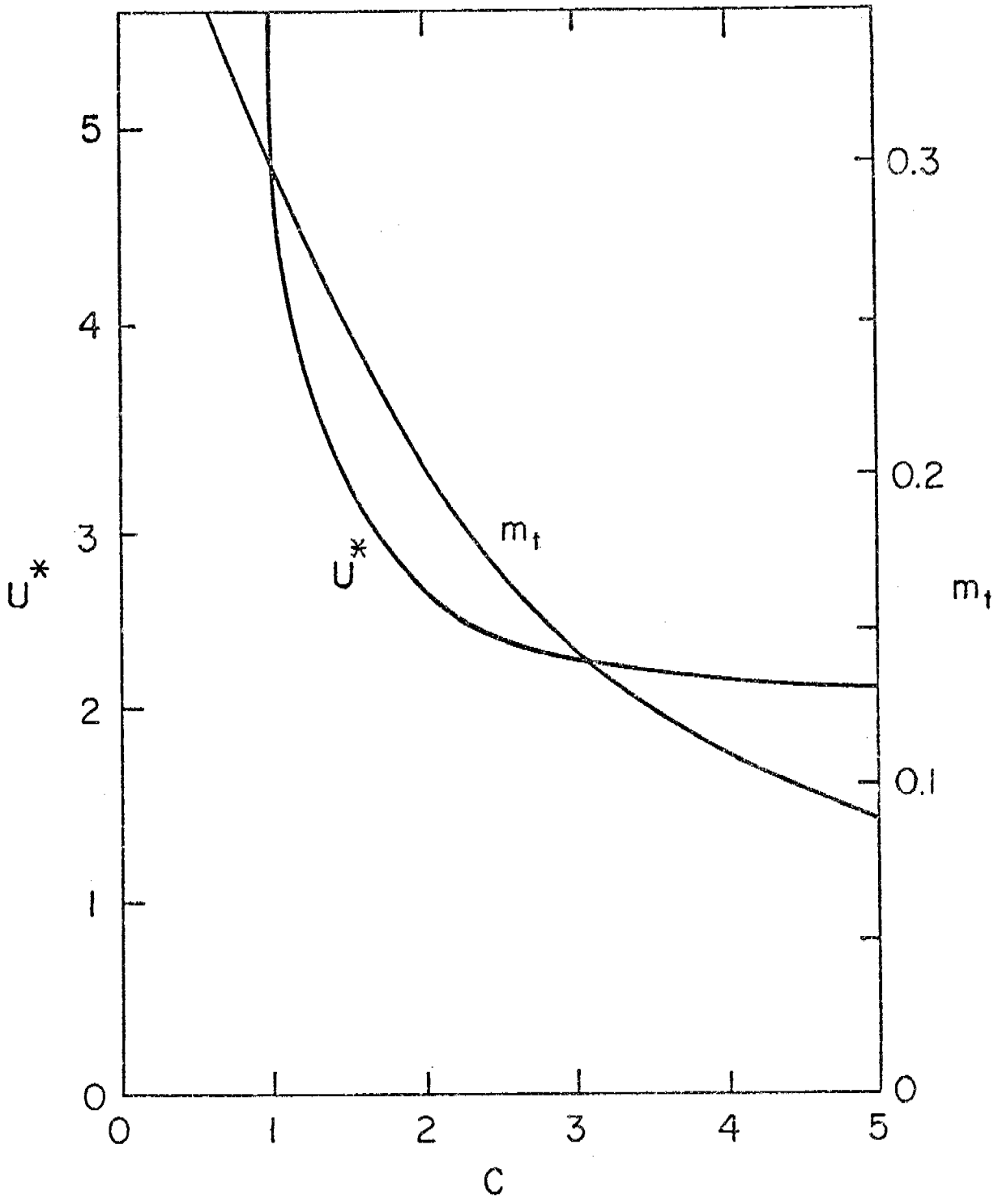


Fig. 6. u^* and m_t Versus c Curves.

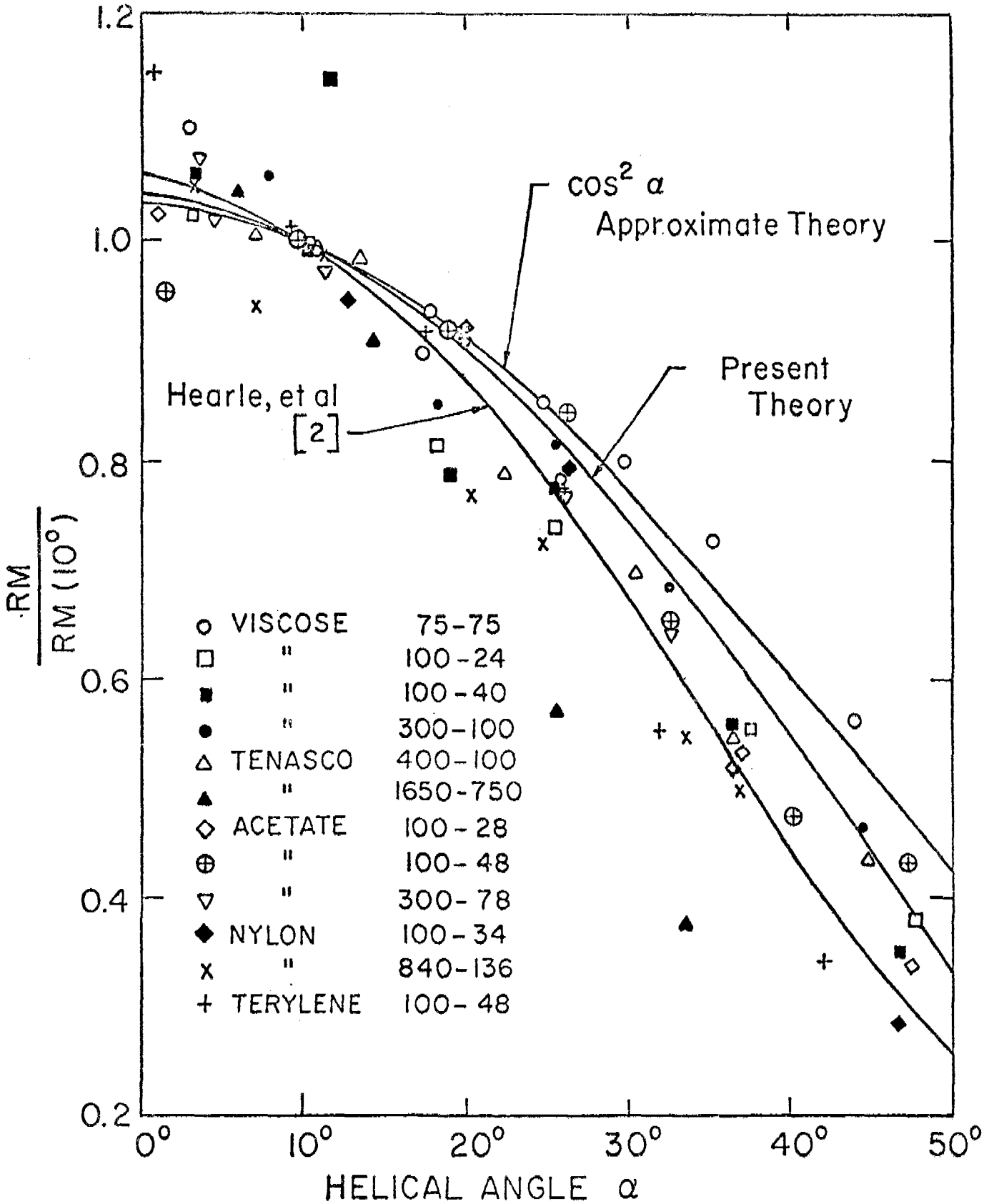


Fig. 7. Comparison of Experimental Values of RM/RM(10°) and the Theoretical Predictions.

obtained by Hearle et al. [3,4] are also shown for comparison. The experimental values are found to be scattered in the vicinity of the theoretical curves as a result of sensitivity of the relative modulus to material properties for different yarns.

The yarn is also sensitive to the twisting moment applied at its ends. It is found that the separated fibers tend to close again as a result of the application of a positive twisting moment of small magnitude. Our analysis can be applied to investigate the behavior of the yarn after all fibers are closed. In Figure 8, the value of U^* and V^* for a yarn with $c = 3$ and closed fibers are plotted against m_t. From those curves, we can draw the same conclusion as what we have made in the problem of the two-ply filament yarns, i.e. the overall axial strain of the yarn is reduced due to the superposition of an additional twisting moment. When m_t is sufficiently large, separation of fibers may occur in the principal normal direction and buckling of the yarn may be introduced. The investigation of the elastic stability of yarns is beyond the scope of this paper.

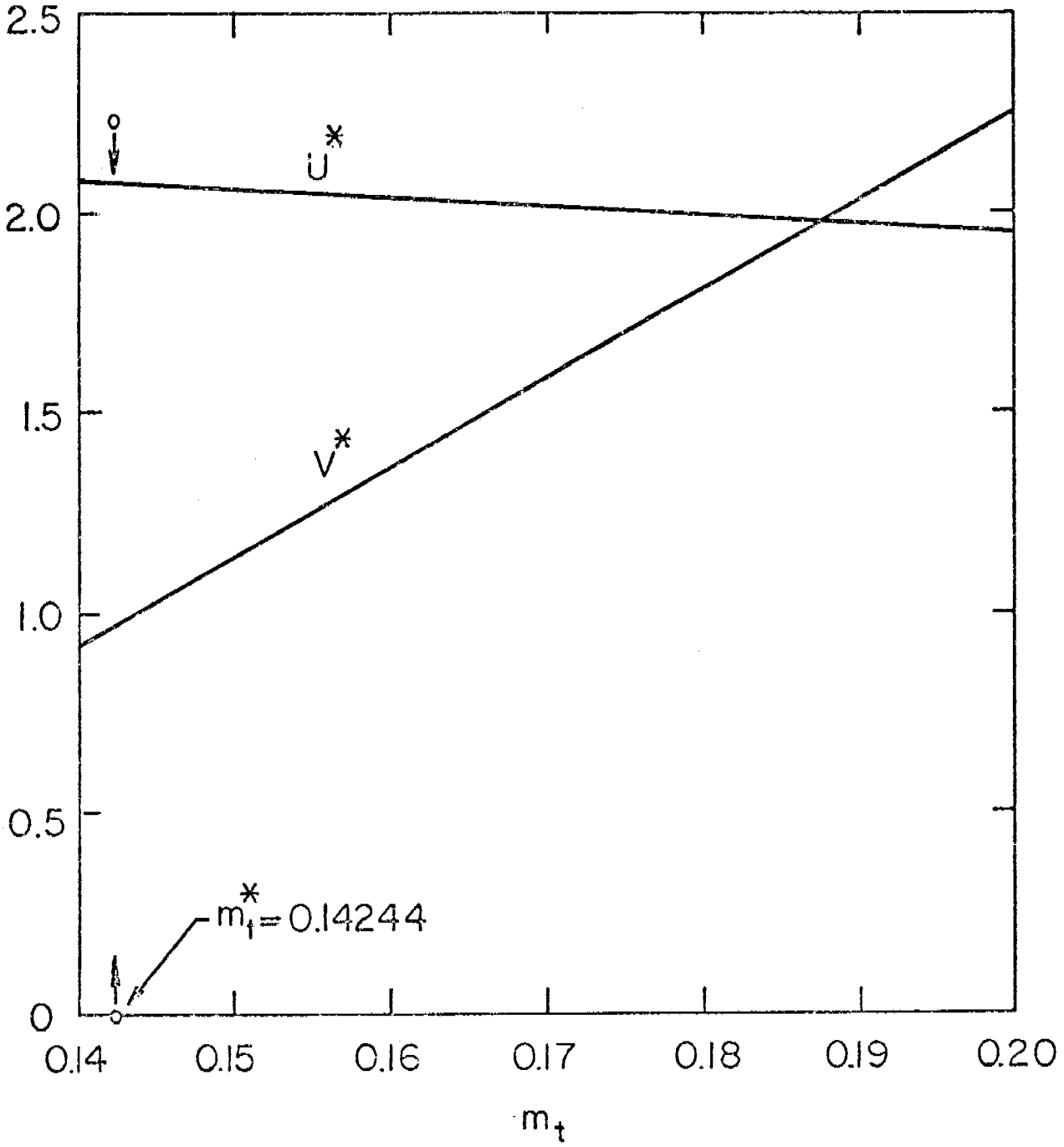


Fig. 8. U^* and V^* Versus m_t Curves.

REFERENCES

1. Wilson, N. and Treloar, L. R. G., "Rubber Models of Yarns and Cords, the 'Doubling' of Single Rods," British Journal of Applied Physics, 12, 1961, pp. 147-154.

2. Hearle, J. W. S., "Theory of the Extension of Continuous Filament Yarns," Structural Mechanics of Fibers, Yarns and Fabrics, edited by J. W. S. Hearle, P. Grosberg, and S. Backer, New York, Wiley-Interscience, 1969, pp. 175-212.

3. Hearle, J. W. S., El-Behery, H. M. A. and Thakur, V. M., "The Mechanics of Twisted Yarns: Theoretical Development," Journal of Textile Institute, 52, 1961, T197-T220.

4. Treloar, L. R. G. and Hearle, J. W. S., "The Mechanics of Twisted Yarns - a Correction," Journal of Textile Institute, 53, 1962, T446-T448.

5. Huang, N. C., "Theory of Elastic Slender Curved Rods," Journal of Applied Mathematics and Physics (ZAMP), 24, 1973, pp. 1-19.

6. Sanders, J. L., Jr., "Nonlinear Theories for Thin Shells," Quarterly of Applied Mathematics, 21, 1963, pp. 21-36.

7. Huang, N. C., "On the Extension of Elastic Two-Ply Filament Yarns," MRC Technical Summary Report #1518, December 1974, University of Wisconsin, Madison, Wisconsin, to appear in Journal of Applied Mechanics.

8. Huang, N. C. and Funk, G. E., "Theory of Extension of Elastic Continuous Filament Yarns," Textile Research Journal, 45, 1975, pp. 14-24.

A MAXIMUM LIKELIHOOD DECISION ALGORITHM FOR MARKOV SEQUENCES WITH MULTIPLE APPLICATIONS TO DIGITAL COMMUNICATIONS

Andrew J. Viterbi
LINKABIT Corporation
San Diego, CA 92121

Abstract

A maximum likelihood decision algorithm is described for Markov sequences and independent observations. At least six different applications of this algorithm to the field of digital communications have been implemented or proposed. The three most important, convolutional coding, intersymbol interference, and voice compression, are summarized and discussed.

1.0 The Basic Algorithm

Decision theory, Bayesian or otherwise, has classically been concerned with the testing of hypotheses involving disjoint events, which are either unique or a sequence of successively independent events. For example, in reliability testing the hypotheses concern the quality of the lot. In radar, it is whether a target is present or absent, and if more than one target is to be tested, successive events are taken to be independent. In block coded digital communications M hypotheses regarding the messages, or codewords of the block code, are to be tested, but successive codewords are assumed to be independent of one another.

On the other hand, there are numerous interesting applications of decision theory for which the hypotheses correspond to the states of a Markov chain. Figure 1.1 represents a fairly general system model of such applications. For a memoryless channel each term of the observation sequence depends only on the corresponding term of the transition sequence. That is, $y(k)$ depends only on $\xi(k)$ and not on previous and successive transitions.

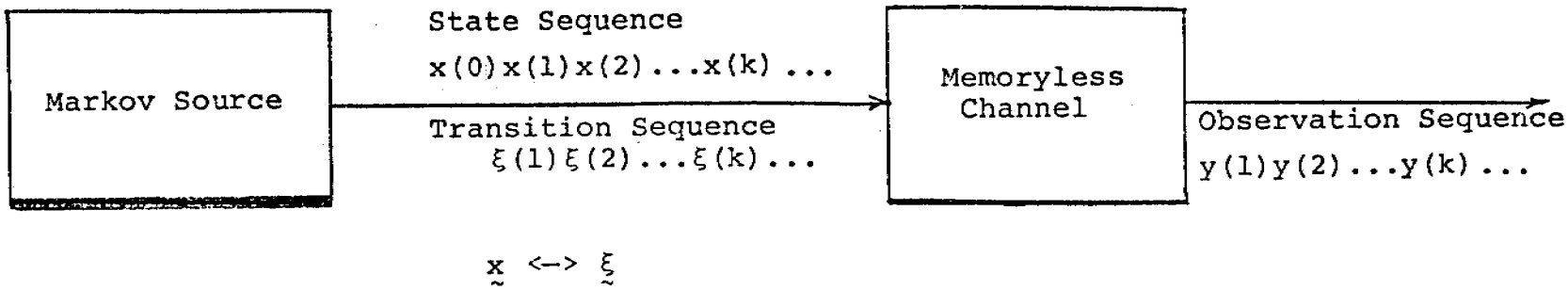
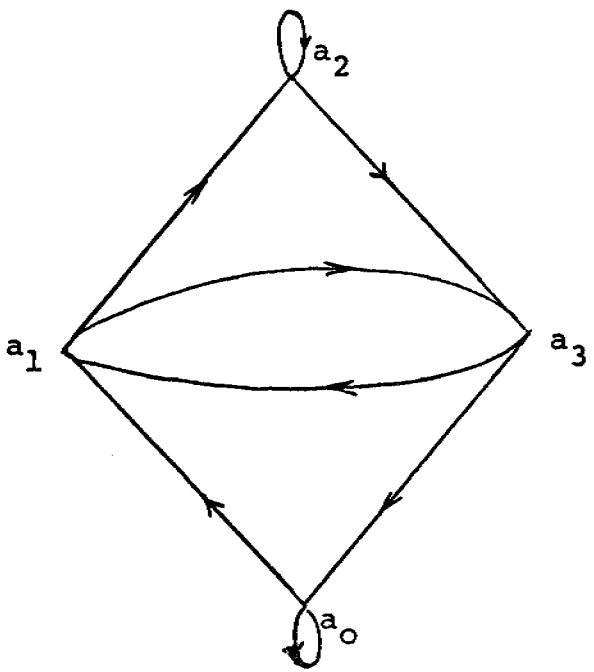


Figure 1.1 General System Model



Initial State $x(0) = a_o$

Final State $x(L) = a_o$

Figure 1.2 State Diagram (only permissible transitions shown)

The Markov source generating the state sequence and hence the transition sequence, which is in one-to-one correspondence with the state sequence, can be depicted by a Markov state diagram as shown in Figure 1.2. Only the permissible transitions are shown; all others are assumed to have transition probability zero. With very little loss of generality, the initial and final states of the source are both taken to be state a_o. In practice even this requirement can be avoided. Superficially the model of Figure 1.1 is similiar to the estimation theoretic model upon which Kalman filtering is based. However, the Markov source here generates a finite state sequence while the estimation theoretic case involves a real sequence.

Table 1.1 contains a derivation of the maximum likelihood or, more precisely, the maximum A Posteriori decision rule. The key expression contained in the box is derived on the basis of three simple observations.

(1) There is a one-to-one correspondence between the state sequence and the transition sequence. Hence, the vector $\underset{\sim}{x}$ can be replaced by $\underset{\sim}{\xi}$ at any point;

(2) the channel is memoryless and hence the conditional probability of $y(k)$ given $\xi(k)$ can be factored out from the expression of the joint conditional probability from 1 to K;

(3) the source is Markov and hence the Kth state depends only on the (K-1)th.

We then seek to maximize the expression within the box to determine the maximum a posteriori probability sequence terminating at state a_j at time K. With the definition of Λ_k and λ_k, it then follows easily that the maximum at time K to state a_j must have been a maximum to some state a_i, $i \in (1, 2, \ldots, J)$, at time K-1. This gives rise to the recursive algorithm shown at the bottom of Table 1.1, where $\Lambda_k(a_j)$ is called the path metric, being proportional to the logarithm of the maximum a posteriori probability for paths terminating in state a_j at time K, and it corresponds to a path which we call $\pi_k(a_j)$. That path must have as its last term the symbol a_j and all previous terms depend on the history of the path; that is, the states at which it resided at previous times. Initial and final conditions are derived from the fact that we have assumed state a_o to be both the initial and final state. Hence the logarithm of the probability of a_o at 0 is 0, and for any other state at time 0, it is $-\infty$. Similarly, the final condition reflects the fact that with probability 1 the final transition must be to state a_o.

Table 1.1 Maximum A Posteriori Decision Rule

State Sequence $\underset{\sim}{x}$ $\xleftarrow{1\text{-}1}$ Transition Sequence $\underset{\sim}{\xi}$ Observation Sequence $\underset{\sim}{y}$: $\xi(k) \to y(k)$

$x(k) \in \{a_o, a_2 \ldots a_J\}$ $\quad x(0) = a_o \quad x(L) = a_o$

$$P[x(1)\ldots x(K)\,|\,y(1)\ldots y(K)] = P[x(1)\ldots x(K),\ y(1)\ldots y(K)]/P[y(1)\ldots y(K)]$$

$$P[y(1)\ldots y(K),\ x(1)\ldots x(K)] = P[y(1)\ldots y(K),\ \xi(1)\ldots\ \xi(K)] = P[y(1)\ldots y(K)\,|\,\xi(1)\ldots\xi(K)]P[x(1)\ldots x(K)]$$

$$= P[y(1)\ldots y(K-1)\,|\,\xi(1)\ldots\xi(K-1)]P[y(K)\,|\,\xi(K)]P[x(1)\ldots x(K-1)]P[x(K)\,|\,x(K-1)]$$

$$\boxed{\log P[y(1)\ldots y(K),\ x(1)\ldots x(K)] = \log P[y(1)\ldots y(K-1),\ x(1)\ldots x(K-1)] + \log\{P[y(K)\,|\,\xi(K)]P[x(K)\,|\,x(K-1)]\}}$$

Define $\Lambda_K(a_j) \overset{\Delta}{=} \underset{\substack{\text{all paths}\\ \text{terminating}\\ \text{at state } a_j\\ \text{at time } K}}{\text{Max}} \log P[y(1)\ldots y(K),\ x(1)\ldots x(K)] \overset{\Delta}{=} \log P[y(1)\ldots y(K),\ \hat{x}(1)\ldots\hat{x}(K-1)a_j]$

$$\lambda_K(a_i \to a_j) \overset{\Delta}{=} \log\{P[y(K)\,|\,\xi(K): a_i \to a_j]P[x(K) = a_j\,|\,x(K-1) = a_i]\}$$

Recursive Algorithm:

$$\Lambda_K(a_j) = \underset{i\in\{1,2,\ldots J\}}{\text{Max}} [\Lambda_{K-1}(a_i) + \lambda_K(a_i \to a_j)]$$

Path Metric for state a_j at time K for path $\pi_K(a_j) \overset{\Delta}{=} \hat{x}(1)\hat{x}(2)\ldots\hat{x}(K-1)a_j$

Initial Conditions:

$\Lambda_o(a_o) = 0; \quad \Lambda_o(a_i) = -\infty \qquad \forall i \neq 0$

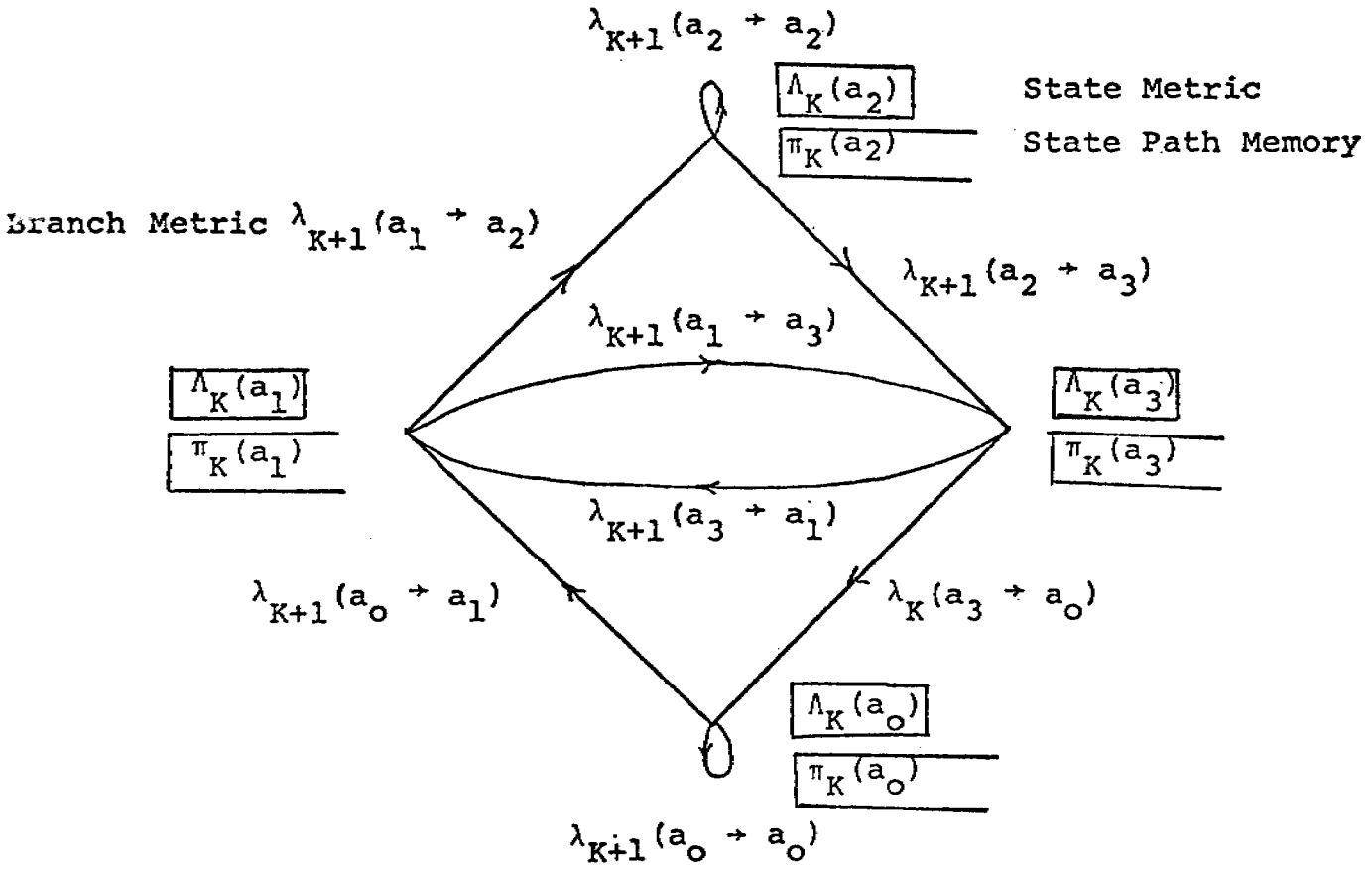
Final Condition:

$\lambda_L(a_i \to a_j) = -\infty \qquad \forall j \neq 0$

A pictorial view of the algorithm, which also represents its mechanization, is shown in Figure 1.3. Both the state metric and state path memory are stored in registers which are continually updated. Thus, at time K+1, for example, the path metric for state a_2 is chosen as the maximum of the path metric at time K for state a_1 incremented by the branch metric for the transition from a_1 to a_2, and of the state metric at time K for state a_2 incremented by the branch metric for the transition from a_2 to itself. Depending on which of these two terms, which we call Γ_1 and Γ_2, is greater, the corresponding state path memory is chosen and incremented by a_2, which corresponds to the state that we are looking at, at time K+1. Thus the registers for the state path memory are shown of indefinite length and they grow linearly with time. Even this requirement can be avoided in practical systems.

We now proceed to describe three major applications of the algorithm.

Figure 1.3 Algorithm Implementation



Example: $\Lambda_{K+1}(a_2) = \text{Max}\ \{[\underbrace{\Lambda_K(a_1) + \lambda_{K+1}(a_1 \to a_2)}_{\Gamma_1}], [\underbrace{\Lambda_K(a_2) + \lambda_{K+1}(a_2 \to a_2)}_{\Gamma_2}]\}$

If $\Gamma_1 > \Gamma_2$, $\pi_{K+1}(a_2) = \pi_K(a_1), a_2$

$\Gamma_2 \leq \Gamma_1$, $\pi_{K+1}(a_2) = \pi_K(a_1), a_2$

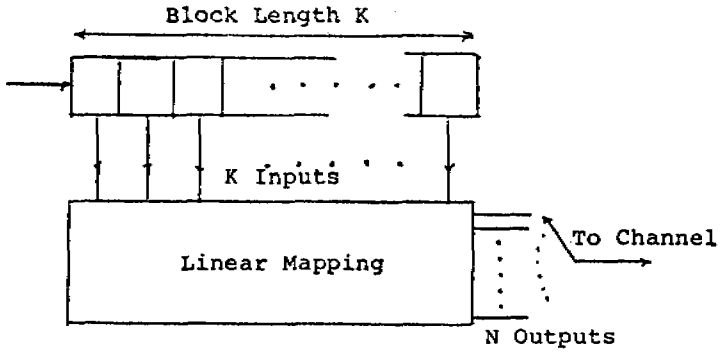
2.0 First Application: Convolutional Codes

The most widespread and significant application of the algorithm just described is for convolutional codes which improve performance on a noisy channel. Superficially a convolutional code generator is very similar to a block code generator, as shown in Figure 2.1. The only difference is that whereas a block code generator takes K successive bits and generates N channel symbols, a convolutional code generator enters one bit at a time into the encoder register and generates n bits where n is on the order of $\frac{N}{K}$ in the block code case. Actually the diagram in Figure 2.1 can be generalized to the case where k > 1 bits are entered at one time, but k is much less than K in most cases of interest.

The convolutional coder can be regarded as a linear finite state machine whose states constitute a Markov sequence. In fact, the Markov graph of Figure 1.2 applies to the convolutional code of constraint length K = 3, and for all convolutional codes of rate 1/n the state or Markov diagram is connected in a binary fashion; that is from each state there are branches going to two states and branches coming to it from two states. As shown at the bottom of Figure 2.1, the state $\underset{\sim}{x}0$ has branches going to $1\underset{\sim}{x}$ and $0\underset{\sim}{x}$ where $\underset{\sim}{x}$ is an arbitrary K-2 dimensional vector; similarly state $\underset{\sim}{x}1$ has branches going out to the same two states.

Figure 2.1 First Application: Convolutional Codes

Block Code: Fill Register with K bits
Generate N Outputs
Rate R = K/N < 1

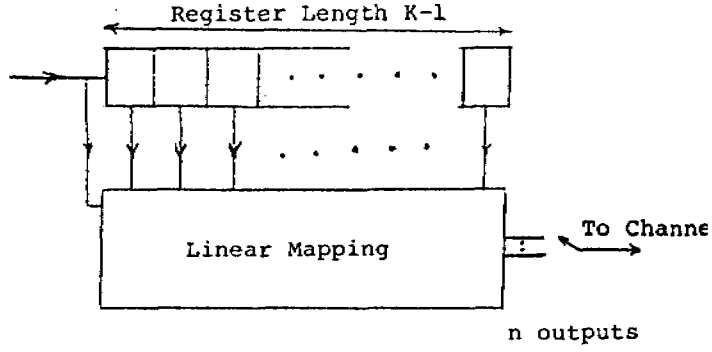


MAP Decision Rule:

Hypotheses <—> Disjoint Events

Classical Solution involves comparison of 2^K likelihood functions or A Posteriori probabilities

Convolutional Code, Contraint Length K,
Input 1 bit at a time
Generate n outputs
R = 1/n



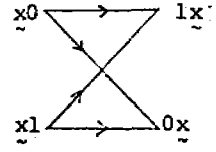
(May generalize to k < K inputs for each n outputs: R = k/n)

MAP Decision Rule:

Hypotheses <—> Markov Sequences

2^{K-1} States correspond to Register Contents

Permissible State Transitions:



e.g. For K = 3, state diagram of Figure 1.2

Table 2.1 represents the key asymptotic results in comparing block and convolutional coding. For all but a very small rate region, the lower bound exponent $\tilde{E}$ is approximately equal to the upper bound exponent E. The most important difference in performance between block and convolutional codes is that whereas the block code exponent E_b is a positive convex function for all rates less than capacity, the convolutional code exponent $E_c(R)$ is concave for most channels, and certainly for all channels of interest.

Figure 2.2, in fact, shows these exponents for a typical channel (e.g., the additive Gaussian channel or the binary symmetric channel). The complexity of the decoder Γ, as already observed and indicated in Figure 1.3, grows as 2^K since this is the storage required as well as the number of computations and comparisons per bit. Similar observations can be made for block codes. Thus the error probability decreases as $\Gamma^{-E(R)/R}$. For this reason convolutional codes outshine block codes of the same complexity.

As a practical example, for the convolutional code of constraint length K = 7 and rate 1/2, the required channel signal-to-noise for a bit error probability of 10^{-5} is reduced by 5 dB compared to that required without

Table 2.1

Performance Comparison

Block Coding Theorem (Shannon et al)

For any memoryless channel and any K, $\exists$ a block code of length K, rate R, and 2^K codewords for which the bit error probability

$$P_B < 2^{-(K/R)\ E_b(R)}$$

where $E_b(R)$ is a positive convex function $\forall$ R < C, channel capacity. Conversely, for every code with these parameters

$$P_B > 2^{-(K/R)[\tilde{E}_b(R) + o(K)]}$$

[For most channels and most rates of interest, $\tilde{E}_b(R) \approx E_b(R)$]

Convolutional Coding Theorem

For any memoryless channel and any K, $\exists$ a convolutional code of constraint length K, rate R, and 2^{K-1} states for which the bit error probability

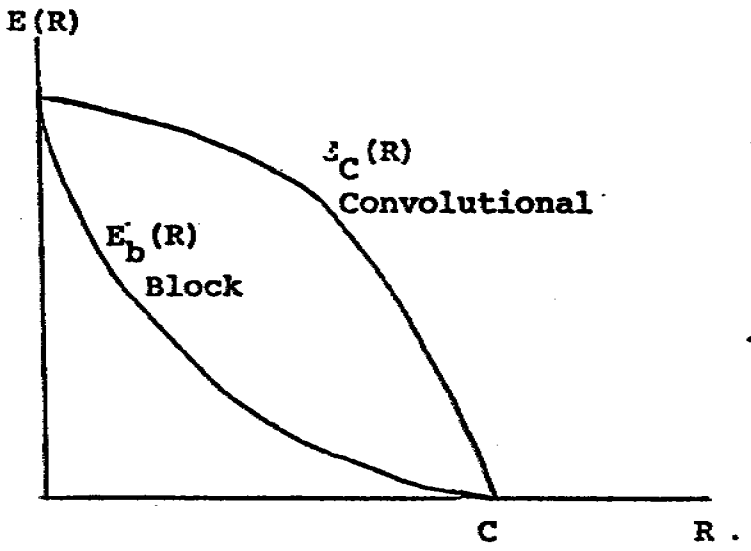
$$P_B < 2^{-(K/R)[E_c(R)-o(K)]}$$ where $E_c(R)$ is a positive function which dominates $E_b(R)$ $\forall$ R < C

For most channels of interest (e.g. symmetric channels) $E_c(R)$ is concave. Conversely, for every code with these parameters

$$P_B > 2^{-(K/R)[\tilde{E}_c(R) + o(K)]}$$

[For most channels and most rates of interest, $\tilde{E}_c(R) \approx E_c(R)$]

For Typical Channel



$P_B \approx 2^{-(K/R)E(R)}$

Decoder Complexity $\Gamma \sim 2^K$

$\therefore P_B \sim \Gamma^{-E(R)/R}$

Practical Result:

For K = 7, Rate 1/2 convolutional code on coherent Gaussian channel, required channel signal-to-noise ratio reduced by 5 dB for $P_B = 10^{-5}$ relative to uncoded operation. Equal complexity block codes gain only 2 to 3 dB over uncoded operation.

Figure 2.2

coding; for block codes of the same complexity, the reduction is only on the order of 2 to 3 dB. Actual hardware with the above parameters and performance, which is presently operational is shown in Figure 2.3. This was developed by LINKABIT Corporation for the U.S. Army Satellite Communication Agency at Ft. Monmouth. Of major significance is the fact that this equipment is capable of decoding at information rates up to 10 megabits per second which are necessary for high capacity and multiple access satellite communications. Figure 2.4 shows the same decoder in the upper right hand corner, with the PSK modem with which it was initially tested. The size of the coder-decoder is approximately one-quarter of that of the modem and yet it gains a 5 dB advantage over the modem alone.

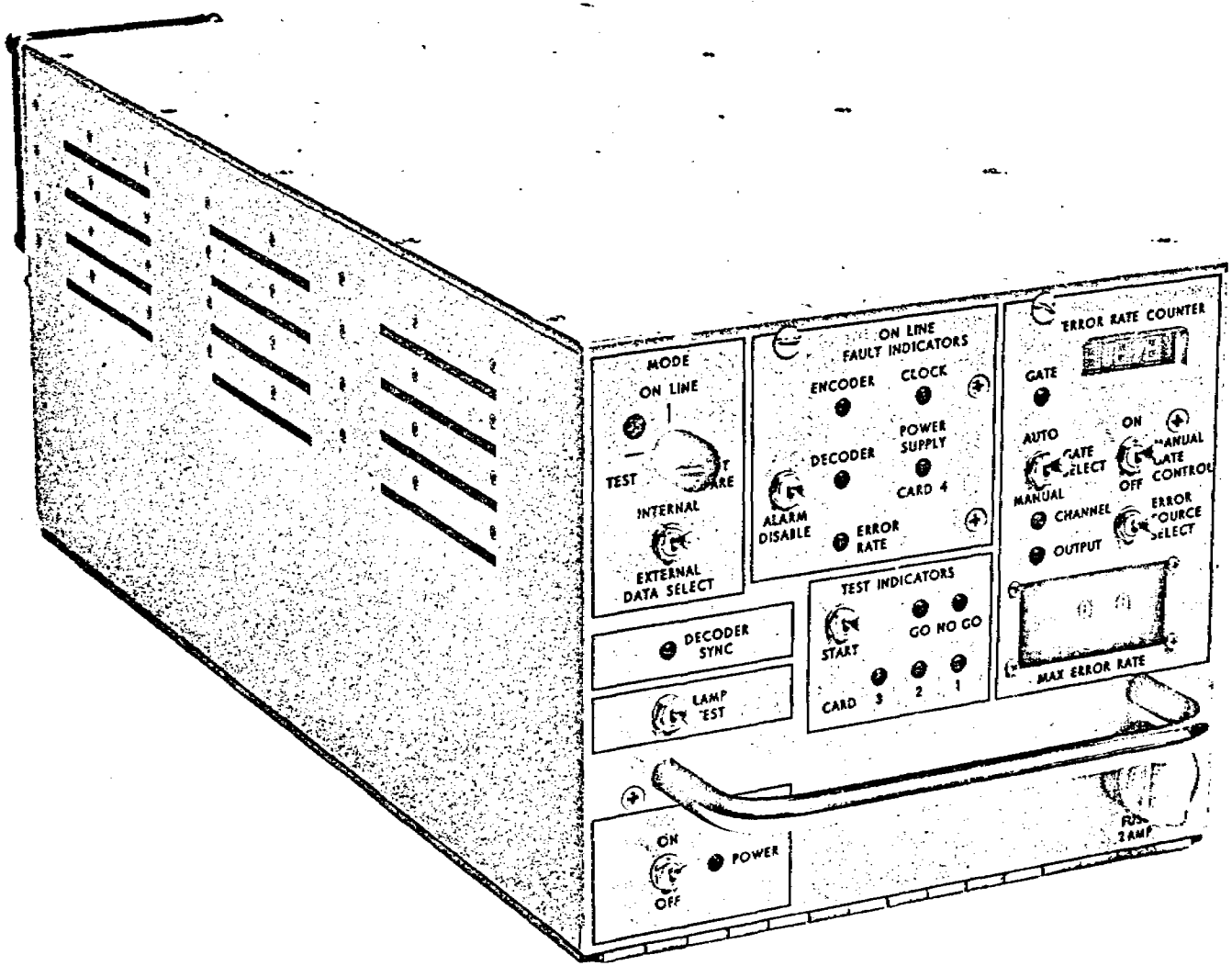


Figure 2.3 LV7017: K = 7, R = 1/2 Convolutional Encoder-Decoder for Information Rates up to 10 Mbps

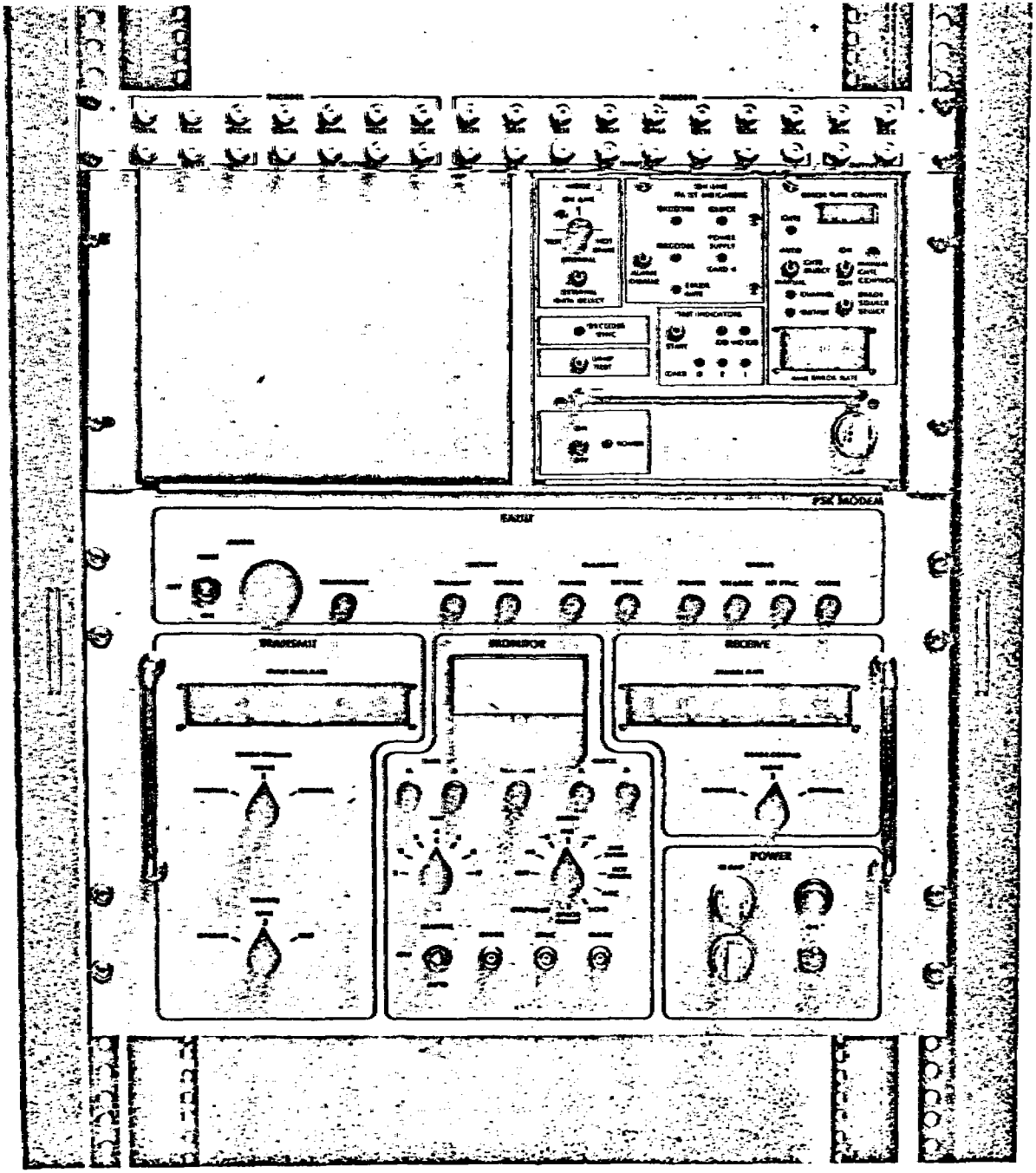


Figure 2.4 LV7017 (upper right corner) Rack Mounted with PSK Modem

3.0 Second Application: Channels with Intersymbol Interference

The second most significant application of the maximum likelihood decision algorithm of Section 1 is to the improvement of performance on channels which exhibit intersymbol interference. Such interference can be caused by either of two physical phenomena illustrated by Figure 3.1:

(a) a bandlimited channel where successive symbols are rendered dependent by the linear filtering characteristics of the channel. The coefficients α_k decay more slowly as the bandwidth of the channel decreases. An example of such a channel results from transmission of digital data over bandlimited telephone lines.

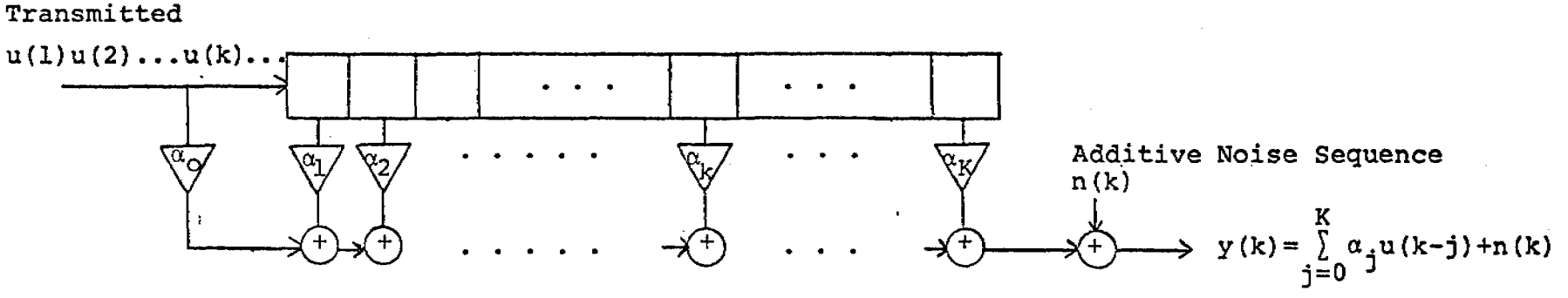
(b) multipath channel, as occurs in HF ionospheric, or tropospheric propagation. Whereas in case (a) the coefficients α_k are usually constant over a message transmission or reasonable portion thereof, in case (b) the coefficients are randomly time varying, although the rate of change is generally low compared to the data rate.

In either case, the coefficients α_k must be estimated either initially or continually in order to establish the model. Rapidly converging estimation algorithms for this purpose are well known and have been thoroughly covered in the literature. Once the model has been established,

Figure 3.1 Second Application: Channels with Intersymbol Interference

Discrete Baseband Model

a) Bandlimited
b) Multipath



a) For Bandlimited Channel $\{\alpha_k\}$ are fixed

b) For Multipath Channel $\{\alpha_k\}$ randomly time varying. Estimation of parameters from observations required - various algorithms with reasonably rapid convergence

Markov Model valid in either case

it is clear that the shift register generates a Markov sequence. The difference between this and a convolutional encoder is that the linear operations are over the real number field rather than over GF(2). However, for a binary information sequence, u(1), u(2),..., u(K), in the register the state of the input is the same as the state of the corresponding input to a binary convolutional coder. Hence, the same algorithm applies to trying to determine what sequence was transmitted, given the received or observed sequence y(K), which exhibits both the effect of the intersymbol interference and additive noise.

Table 3.1 illustrates the key theorem regarding the performance of the algorithm for intersymbol interference channels. P_o is the error probability for the simple memoryless additive white Gaussian channel, while P_I is the probability with intersymbol interference. The coefficients K_u and K_ℓ are relatively small constants, independent of the channel energy-to-noise ε/N_o, and generally within an order of magnitude of each other. For $\gamma = 1$, the performance would be almost the same as without intersymbol interference. In all cases, of course, $\gamma \leq 1$, but the key result is that for K = 1 (i.e. the two tap case), $\gamma = 1$. Hence, there is little or no degradation when the multipath channel consists of two

Table 3.1 Performance of Algorithm for Intersymbol Interference Channels

Theorem (Forney): The Maximum A Posteriori Decision Algorithm applied to a two-level sequence received over channels with intersymbol interference yields a symbol error probability

$$K_\ell P_o(\gamma\varepsilon/N_o) < P_I < K_u P_o(\gamma\varepsilon/N_o)$$

where K_ℓ and K_u are independent of the channel signal-to-noise ratio (ε/N_o), (and are generally within an order of magnitude of each other)

$P_o(\varepsilon/N_o)$ is the error probability without intersymbol interference and $\gamma \leq 1$.

For K = 1 (2 tap case), $\gamma = 1$

(Generalizes to m-level input sequence)

Practical Result: Transmission over multipath channel with only 2 strong paths can be received with very little degradation relative to single path case. (Applies also to data modems employing "duo binary" or "partial response" coding).

strong paths and all other paths are negligible. For telephone channels such a case is often artificially generated by applying so-called duo-binary or partial response coding, which creates a model consisting of only two strong taps.

These results can be generalized to an m-level input sequence. While no operational hardware exists which applies these techniques, considerable experimental work and simulations have been performed.

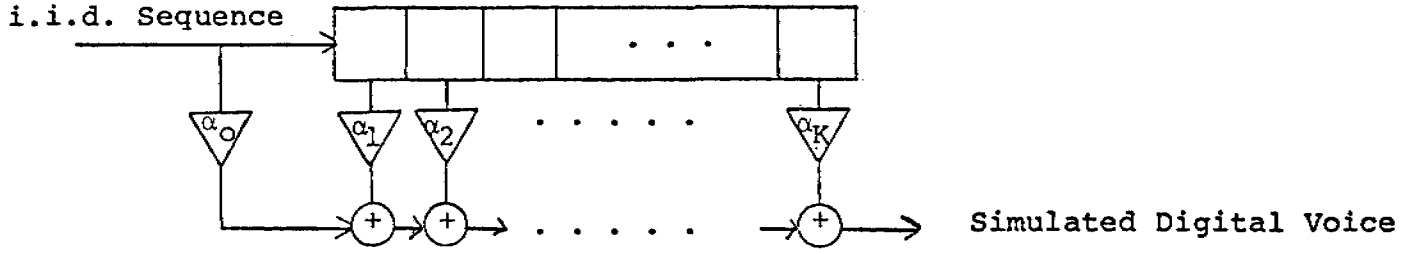
4.0 Third Application: Digital Voice Compression

Potentially the most important application for the future is to the compression of digital voice. In Figure 4.1 the upper diagram shows the well accepted voice synthesizer model used in conjunction with linear predictive coding. This is basically a modern digital version of the model for the channel vocoder. The vocal tract is modelled by a digital linear filter driven either by an i.i.d. sequence of Gaussian variables or by a periodic sequence, the latter representing voiced sounds and the former unvoiced. Clearly then the output is a Markov sequence, although not a finite-state one.

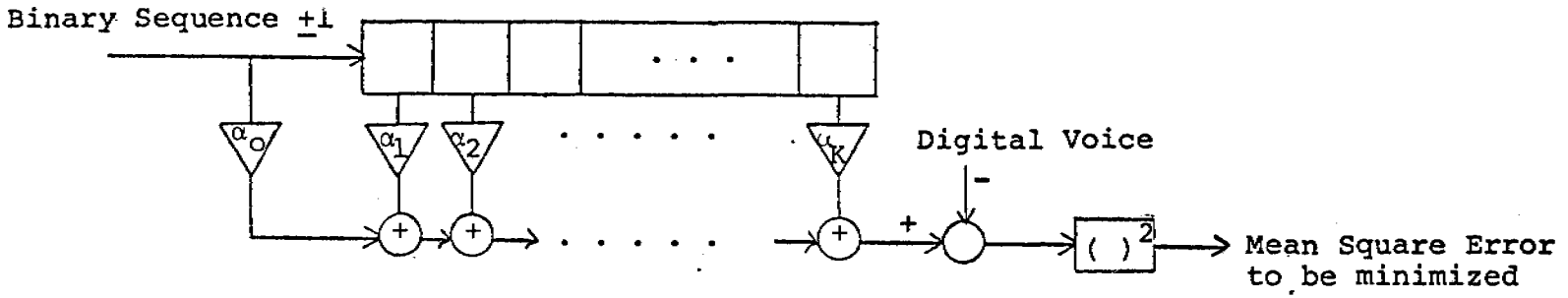
The lower half of the diagram is representative of a large class of standard practical voice compression techniques, including delta modulation, differential PCM, and adaptive predictive coding. In the last, as in linear predictive coding, the coefficients α_o through α_k are estimated from short segments of the actual voice. However, all these techniques attempt to represent the voice by means of the so-called binary residual sequence consisting of $\pm$ 1's. All the conventional techniques make decisions to generate this sequence one term at a time, seeking to minimize the mean square error for that term only. An improved technique for which simulations and digital voice

Figure 4.1 Third Application: Digital Voice Compression

Voice Synthesizer Model (Linear Predictive Coding) - $\{\alpha_k\}$ must be estimated frequently (every 10 to 20 millisec)



Standard Technique for Delta Modulation; Differential PCM; Adaptive Predictive Coding



Select binary sequence one term at a time to minimize MSE for that term only

Improved Technique: Select best path through Markov state diagram generated by linear filter driven by binary sequence to minimize MSE for entire path

Performance and Practical Results: Subjective

compression experiments have been performed at LINKABIT Corporation involves estimating the entire sequence using the algorithm of Section 1. One may regard the digitized uncompressed voice as the observation and one then looks for the binary sequence (i.e., the path through the Markov chain) which best replicates the observation, given, of course, the model and its coefficients. The compressed digital voice consists of the residual sequence plus the digitized estimated coefficients $\{\alpha_k\}$. The voice decoder at the receiver reconstitutes the voice sequence by passing the residuals through the digital linear filter of Figure 4.1 with these tap coefficients.

Evaluation of this technique and any practical application thereof must be based on subjective results. During the lecture a tape was played giving the relative performance of the algorithm based on estimating the Markov sequence as compared to the results of a more classical adaptive predictive coding technique where each term of the residual was determined individually. In addition, the result of adding errors caused by channel noise was shown to affect the algorithm very little as long as the channel error rate was below 10^{-2}.

5.0 Other Applications

Another area of application described in the literature has been optical character recognition, where English text was modelled by a first-order Markov source, and the memoryless channel was represented by the optical character recognition device. The algorithm has also been applied to the demodulation of minimum shift keyed (MSK) modulation and to certain phase tracking problems. Any application which can be modelled as shown in Figure 1.1 is suited to the maximum likelihood decision algorithm described here. The breadth of its applicability is remarkably wide, apparently spanning the gamut of modern digital communications, which includes modems, channel coding and data compression.

THE CALCULATION OF UNSTEADY SHEAR STRESSES IN GUN BARRELS

R. Yalamanchili

Research Directorate
GEN Thomas J. Rodman Laboratory
Rock Island Arsenal
Rock Island, Illinois 61201

ABSTRACT

Since the shear stresses at the wall are of primary interest in gas dynamic erosion, an analytical technique is formulated for unsteady shear forces by consideration of the continuity and momentum equations of unsteady compressible boundary layers. The continuity and momentum equations are transformed into one governing equation by the use of equation of state, power law velocity and temperature profiles, and generalized Blasius' law of friction. The single partial differential equation is integrated across the boundary layer to reduce by one the number of independent variables (three). The resulting hyperbolic-type partial differential equation is transformed into two ordinary differential equations by the characteristics method. These equations are integrated to predict the results in a 30mm weapon system. The results include not only various boundary layer thickness parameters but also the shear stress at the wall or skin friction coefficient which is large near the breech and decreases in the flow direction except near the projectile. The skin friction coefficient decreases with increase in time. The shear stress at the wall is of the order of 100 psi.

INTRODUCTION

The shear forces are not only important in the calculation of drag but also in the formulation of force balance in erosion models. The shear stresses also play an important role in predicting the heat transfer provided that a valid analogy between skin friction and heat transfer is available. The shear stresses acting on the bore surface are essential to formulate and predict the gas dynamic erosion in gun barrels. There are now available numerous empirical or semiempirical methods which provide reliable estimates of turbulent skin friction for zero pressure gradient compressible and incompressible turbulent boundary layer flows. However, this is not the case for unsteady boundary layers. The state-of-the-art in unsteady boundary layers and turbulence models is quite limited. Therefore, any development of analytical techniques with reasonable assumptions is of considerable importance to engineers and designers. A symposium[1] was held on unsteady boundary layers at Laval University. There are also several papers, scattered in the literature, concentrating on numerical solutions[2] of unsteady boundary layer equations. The similarity methods do not play an important role due to severe restrictions on the generality of the flow field. Apparently, there is no working tool yet available to analyze especially the case of unsteady compressible flow with arbitrary pressure gradients and free stream conditions outside the boundary layer.

Yalamanchili presented the Rayleigh-Blasius incompressible flow[3], shock-induced boundary layers[4], and projectile-induced turbulent boundary layers[5] in AIAA, TICOM, and ASME meetings, respectively. Even though the continuity, momentum, and energy equations are considered in the analysis of unsteady turbulent boundary layers[5], the density variation

across the boundary layer is not considered due to tedious derivations. However, another treatment of unsteady boundary layers is considered. The transverse velocity component can be expressed in terms of longitudinal velocity component and density by the use of the continuity equation. The density is expressed in terms of pressure and temperature by the equation of state. Of course, the pressure is a given quantity in any boundary layer where there is no interaction between the viscous and inviscid flows. The temperature is written as a function of the velocity boundary layer thickness parameter and a nearly constant coefficient. Thus, the energy equation is not necessary to close the system of equations. It is to be noted that the temperature is not of prime interest here even though it can be approximated from the assumed functional relationship. Nikuradse[6] experimental data suggests power law velocity profiles for the longitudinal velocity component. The single partial differential equation is integrated across the boundary layer to reduce by one the number of independent variables. The generalized Blasius' law of friction, relating the shear stress at the wall and the longitudinal velocity distribution, is utilized. The resulting hyperbolic type partial differential equation is transformed into two ordinary differential equations by the characteristics method. These equations are integrated to predict the results in a 30mm weapon system. The results include not only various boundary layer thickness parameters but also the shear stress at the wall or skin friction coefficient.

II GOVERNING EQUATIONS

It is convenient to use integral boundary layer equations to apply the method of weighted residuals (MWR). The following equation can be derived either by consideration of control volume and setting continuity (mass) and momentum balance for the same or by application of Prandtl's boundary layer assumptions to the Navier-Stokes equations and integrating the resulting equations across the boundary layer:

Momemtum Equation:

$$-\frac{\partial p}{\partial x}\int dy - \tau_w = \frac{\partial}{\partial t}\int \rho u dy + \frac{\partial}{\partial x}\int \rho u^2 dy - u_1\frac{\partial}{\partial t}\int \rho dy - u_1\frac{\partial}{\partial x}\int \rho u dy \qquad (1)$$

Where u = velocity component in x-direction, u_1 = velocity outside the boundary layer, p = pressure, ρ = density, t = time and τ_w = shear stress at the wall.
Utilizing the following inviscid momentum equation

$$-\frac{\partial p}{\partial x} = \rho\frac{\partial u_1}{\partial t} + \rho u_1\frac{\partial u_1}{\partial x} \qquad (2)$$

and the following identities

$$\frac{\partial}{\partial t}(\rho u_1) = u_1\frac{\partial \rho}{\partial t} + \rho\frac{\partial u_1}{\partial t}$$

$$\frac{\partial}{\partial x}u_1\int \rho u_1 dy = u_1\frac{\partial}{\partial x}\int \rho u dy + \frac{\partial u_1}{\partial x}\int \rho u dy \qquad (3)$$

the following single governing partial differential equation is obtained.

$$\tau_w = \frac{\partial}{\partial t}\int (\rho_1 u_1 - \rho u)\, dy + \frac{\partial}{\partial x}\int \rho u(u_1 - u)dy + \frac{\partial u_1}{\partial x}\int (\rho_1 u_1 - \rho u)dy$$

$$- u_1\frac{\partial}{\partial t}\int (\rho_1 - \rho)\, dy \qquad (4)$$

Nikuradse[6] measurements of velocity distribution in pipes suggested the following form:

$$\frac{u}{u_1} = \left(\frac{y}{\delta}\right)^{1/n} \tag{5}$$

Where δ is the velocity boundary layer thickness, function of independent variables x and t, and n is a parameter depending upon the Reynolds number based upon mean velocity and diameter of the tube:

n	6	7	8	8.8	10
Reynolds number	4×10^3	2.3×10^4	1.1×10^5	1.1×10^6	2 to 3.2 millions

The equation of state and the fact that the pressure is constant across the boundary layer yields the following relationship for density:

$$\frac{\rho}{\rho_1} = \frac{T_1}{T} \tag{6}$$

Where T is the temperature and the subscript 1 indicates the quantity outside of the boundary layer. If another power law profile is assumed for the dimensionless temperature (i.e., equation 5); and setting T_w equals zero without any loss of generality, the density ratio (equation 6) may be rewritten as equation (8).

$$\frac{T-T_w}{T_1-T_w} = \left(\frac{y}{\zeta}\right)^{1/m} \tag{7}$$

$$\frac{\rho}{\rho_1} = C \left(\frac{y}{\delta}\right)^{-1/m} \quad \text{where } C = \left(\frac{\zeta}{\delta}\right)^{1/m} \tag{8}$$

The variable ζ represents the temperature boundary layer thickness and it is also greater than the velocity boundary layer thickness δ for the conditions in a gun barrel. Since the dimensionless ratio of boundary layer thicknesses (ζ/δ) is fairly constant with a typical value of 2.5 and the exponent m is large similar to n, it is a good approximation to treat C as independent of x and t.

Prandtl discovered the relationship between Blasius' law of friction and the velocity distribution due to Nikuradse experimental data. If this relationship is generalized to include arbitrary n, the following result is obtained for shear stress at the wall.

$$\tau_w = \frac{C_f}{2} \rho_1 u_1^2 = \frac{C_n}{\left(\frac{u_1 \delta}{\nu}\right)^{\frac{2}{1+n}}} \rho_1 u_1^2 \qquad (9)$$

Where C_f is the skin friction coefficient and C_n is given in the form of a table:

n	7	8	9	10
C_n	.0228	.0174	.0143	.0117

The use of equations (5), (8), and (9); the evaluation of integrals in equation (4), and algebraic manipulations will yield the following hyperbolic type of partial differential equation with only two independent variables, x and t.

$$P \frac{\partial \delta}{\partial x} + \frac{Q}{u_1} \frac{\partial \delta}{\partial t} = R\,(\delta, x, t) \qquad (10)$$

where $R = A C_f - \frac{B}{u_1} \frac{\partial}{\partial t} \ell n\, \rho_1 - \frac{C}{u_1} \frac{\partial}{\partial t} \ell n\, u_1 - D \frac{\partial}{\partial x} \ell n\, u_1 - E \frac{\partial}{\partial x} \ell n\, (\rho_1 u_1^2)$ (11)

The coefficients P,Q,A,B,c,D, and E are given in the following table for compressible flow and the case of constant density across the boundary layer.

Coefficient	Compressible flow $\rho(x,y,t)$	Constant density across the boundary layer, $\rho(x,t)$
P	$\frac{n}{2m-n+mn}$	$\frac{n}{n+2}$
Q	$\frac{1}{m-1}$	1
A	$\frac{m-n+mn}{2Cm^2}$	$\frac{n+1}{2}$
B	$\frac{1}{m-1}\ \delta$	$\frac{1}{n+1}\ \delta$
c	$\frac{m-n+mn(1-C)}{Cm^2}\ \delta$	$\frac{1}{n+1}\ \delta$
D	$\frac{m-n+mn(1-C)}{Cm^2}\ \delta$	$\frac{1}{n+1}\ \delta$
E	$\frac{n}{2m-n+mn}\ \delta$	$\frac{n}{(n+1)(n+2)}\ \delta$

It is clear from this table that the case of constant density across the boundary layer is not easy to derive from the general case of compressible flow. The following table is given to show the effect of variable density on coefficients.

The Constant in Coefficient	m = 6, n = 8	
	Compressible flow	Constant density across the viscous layer
P	.15	0.80
Q	.20	1.00
A	.64	4.50
B	.20	0.11
c	-.06	0.11
D	-.06	0.11
E	.15	0.09

Some coefficients decrease whereas others increase and the changes are not proportional. Therefore, it is difficult to assess the overall effect of variable density across the boundary layer.

The simplified governing equation (10) can be reduced to two ordinary differential equations by a procedure sometimes referred to as the method of characteristics. The perfect or total differential can be written as

$$d\delta = \frac{\partial \delta}{\partial x} dx + \frac{\partial \delta}{\partial t} dt \qquad (12)$$

The substitution of the partial derivative $\partial\delta/\partial t$ from equation (10) into equation (12) yields the following:

$$d\delta = (dx - \frac{Pu_1}{Q} dt) \frac{\partial \delta}{\partial x} + \frac{u_1 R}{Q} dt \qquad (13)$$

If the quantity inside the paranthesis can be set to zero and these resulting curves considered as velocity characteristics, equation (13) can be reduced to

$$\frac{d\delta}{dt} = \frac{u_1 R}{Q} = f(\delta, x, t) \tag{14}$$

along the velocity characteristic

$$\frac{dx}{dt} = \frac{u_1 P}{Q} \tag{15}$$

Note that the same equations (14) and (15) can also be derived by following the classical definition of characteristics. The characteristics are defined as the curves along which the derivatives of the fluid properties, such as $\partial\delta/\partial x$ and $\partial\delta/\partial t$, are indeterminate. By consideration of these two partial derivatives as unknowns in equations (10) and (12), one can determine either one of these two unknowns by Cramer's rule and set its numerator and denominator to zero in order to obtain the equations (14) and (15), respectively.

III BOUNDARY LAYER PARAMETERS

The effects of the viscous shear layer are not only induction of shear stresses on the wall but also involve various effects on the gas flow. These effects can be represented by means of various boundary layer parameters. For example, the displacement thickness (δ_d) is defined as

$$\delta_d = \int_0^\delta \left(1 - \frac{\rho u}{\rho_1 u_1}\right) dy \tag{16}$$

which represents physically the distance by which the external inviscid flow is shifted owing to the formation of the boundary layer. The momentum thickness is defined as

$$\delta_m = \int_0^\delta \frac{\rho}{\rho_1} \frac{u}{u_1} \left(1 - \frac{u}{u_1}\right) dy \tag{17}$$

This parameter is useful in the determination of laminar-turbulent transition and also indicates a measure of loss of momentum in the boundary layer. The velocity thickness is defined as

$$\delta_v = \int_0^{\delta} \left(1 - \frac{u}{u_1}\right) dy \tag{18}$$

The energy dissipation thickness is defined as

$$\delta_{ed} = \int_0^{\delta} \frac{\rho}{\rho_1} \frac{u}{u_1} \left(1 - \frac{u}{u_1}\right)^2 dy \tag{19}$$

This parameter indicates a loss of mechanical energy occurring in the boundary layer. The enthalpy thickness is defined as

$$\delta_e = \int_0^{\delta} \frac{\rho}{\rho_1} \frac{u}{u_1} \left(\frac{T}{T_1} - 1\right) dy \tag{20}$$

All of these parameters can be evaluated analytically by the use of equations (5), (6), and (8). The dimensionless parameters (δ_d/δ, δ_m/δ, etc.), are shown in the following table for the case of compressible flow as well as constant density across the boundary layer.

Dimensionless Parameter	Compressible flow	Constant density across the boundary layer
Displacement thickness	$\frac{m-n+mn(1-C)}{m-n+mn}$	$\frac{1}{n+1}$
Momentum thickness	$\frac{Cm^2n}{(m-n+mn)(2m-n+mn)}$	$\frac{n}{(n+1)(n+2)}$
Velocity thickness	$\frac{1}{n+1}$	$\frac{1}{n+1}$
Energy Dissipation thickness	$\frac{2Cm^3n}{(m-n+mn)(2m-n+mn)(3m-n+mn)}$	$\frac{2Cn}{(m+1)(n+2)(n+3)}$
Enthalpy thickness	$n(\frac{1}{n+1} - \frac{Cm}{m-n+mn})$	0

All of these parameters will be larger for the case of variable density than constant density across the boundary layer.

IV GUN BARREL FLOW

The physical example is represented schematically as shown in Figure 1. As the propellant gases expand behind the projectile, a boundary layer forms at the breech end and thickens as the flow proceeds downstream. An unusual feature of the velocity boundary layer is that it disappears as the projectile is approached since all fluid at the base of the projectile must be moving at projectile velocity. Mathematically, this amounts to the requirement of an additional boundary condition at a downstream location. The numerical techniques applied to most boundary layer problems call for the specification of profiles at the upstream end of the flow and allow a "marching"

along the flow direction. For the usual time-dependent boundary layer problem, an initial condition to describe the boundary layer flow at time zero and boundary conditions as functions of time are required. No downstream condition is added. The complete flow characteristics are unknown for gun barrel flows. The experimental data are lacking because of the moving projectile. This obstacle may be overcome if one takes advantage of the similarities between the moving projectile (small mass) and a moving shock in a shock tube. Typical velocity profiles are also shown in Figure 1.

Fortunately, the final governing equations (14) and (15) are decoupled. Therefore, one can solve these equations one at a time. The path of the velocity characteristics are described by a linear ordinary differential equation (equation 15). If the LaGrangian approximation of interior ballistics is invoked here to describe the free stream velocity, u_1, the integration of equation (15) yields equation (17).

$$u_1 = \frac{x}{X(t)} V(t) \tag{16}$$

$$\left(\frac{x}{x_i}\right) = \left(\frac{X}{X_i}\right)^{\frac{n(m-1)}{2m-n+mn}} \tag{17}$$

where V = Velocity of the projectile

X = Location of the projectile

and subscript i indicates characteristic location initially. For the characteristics located at the base of the projectile, x_i and X_i will be identical.

Since the location of characteristics at any time are known from equation (17), the nonlinear ordinary differential equation (14) is integrated numerically by iterative technique due to lack of analytical possibility. The following steps are in order ($d\delta/dt = f$):

- All values are known at point i-1 (could be initial or previous time).
- The values except the dependent variable are known at point i. An approximate value can be obtained by computing f at point i-1, i.e., f_{i-1} and the use of the relationship, $\delta_i = \delta_{i-1} + f_{i-1}\,\Delta t$.
- Compute now f_i with the above approximate dependent variable value. Finally, compute $\delta_i = \delta_{i-1} + \Delta t\,(f_i + f_{i-1})/2$.
- Repeat the above (last) step until the desired degree of convergency is reached. The convergence to a final value is very rapid. It is found that no more than three iterations are required.

This procedure is repeated for each time step and for all characteristics. The CPU time on IBM 360/65 is of the order of 5 seconds for 11 characteristics and 16 time steps.

Since each characteristic curve can be calculated without reference to adjoining characteristics, the accuracy of the calculations for velocity boundary layer thickness, and in turn various boundary layer parameters including the skin friction coefficient does not depend on the number of characteristics chosen. A large number of characteristics mean the results are given at closer intervals in the stream-wise direction, but these are not more accurate at a calculated point. A large number of characteristics will lead to more computational times.

The accuracy of the integration of the boundary layer along each characteristic depends on the time interval chosen. Accuracy and computational times increase with decrease in the time step. Because of the iteration technique used in each step, very small times are not required for good accuracy.

There is a singularity problem in the initiation of calculations. Since δ and V become zero initially (t=o) in the important terms of the right hand side of equation (14) or in the following equation, the integration procedure described above fails.

$$\frac{d\delta}{dt} = (m-1)\left[\frac{2AC_n u_1}{\left(\frac{u_1\delta}{\nu}\right)^{2/1+n}} - \frac{B}{\rho_1}\frac{\partial \rho_1}{\partial t} - c\left(\frac{1}{V}\frac{dV}{dt} - \frac{1}{X}\right) - \frac{(D+2E)u_1}{x} - \frac{Eu_1}{\rho_1}\frac{\partial \rho_1}{\partial x}\right] \qquad (18)$$

This is due to equation (16) and its variations:

$$\frac{\partial}{\partial t} \ell n\, u_1 = \frac{1}{V}\frac{dV}{dt} - \frac{1}{X} \qquad (19)$$

$$\frac{\partial}{\partial x} \ell n\, u_1 = \frac{1}{x} \qquad (20)$$

To initiate the computations when δ and V become zero, equation (18) is approximated to

$$\frac{d\delta}{dt} = (m-1)\left[\frac{2AC_n u_1}{\left(\frac{u_1\delta}{\nu}\right)^{2/1+n}} - \frac{c}{u_1}\frac{\partial u_1}{\partial t}\right] \qquad (21)$$

Taking $u_1 = A_1 t$, where A_1 is the local gas acceleration at position x and constant for the first time step, equation (21) becomes

$$\frac{d\delta}{dt} = \frac{m-1}{Cm^2}\left[\frac{C_n(m-n+mn)u_1}{\left(\frac{u_1\delta}{\nu}\right)^{2/1+n}} - \left(\frac{\delta}{t}\right)(m-n+mn(1-C)\right] \qquad (22)$$

$$\begin{aligned} \text{Let} \quad \phi &= 1+n \\ \delta &= \psi^{1/\phi} \\ p &= \frac{m-1}{Cm^2}(1+\phi)(m-n+mn(1-C)) \\ q &= \frac{m-1}{Cm^2} Cn\, \nu^{\phi} A_1^{1-\phi} (1+\phi) \end{aligned} \qquad (23)$$

then, equation (22) becomes

$$\frac{d\psi}{dt} + \frac{p}{t}\psi = qt^{1-\phi} \qquad (24)$$

The solution of equation (24) is

$$\psi = \frac{q}{2+p-\phi} t^{2-\phi} \qquad (25)$$

or $$\delta = \frac{(m-1)(2+n)C_n \nu \phi A_1^{-n} t^{1-n}}{Cm^2(1-n) + (m-1)(2+n)(m-n+mn(1-C))} \quad (26)$$

This is used to initiate the computations for the first two time steps in order to avoid the singularity problem. For the projectile initiated characteristics (i.e., at the projectile base) where only δ becomes zero but not V, equation (14) or (18) can be multiplied by $\delta^{2/\phi}$ and the resulting equation can be integrated numerically as mentioned above without any problem of singularity.

The following data is assumed for the XM140 (30mm) weapon system in order to compute the characteristics of velocity boundary layers:

Barrel Length = 42.0 inches
Chamber Length = 2.48 inches
Reference dynamic viscosity at 530°R = 0.00003 lbs/ft-sec
m = 7, n = 7
time step = 0.0001 seconds

Time	Density lb f/ft^3	Projectile Velocity (ft/sec)	Temperature,°R
0.00000	6.5	0	5200
0.00040	8.2	-	-
0.00085	5.0	1600	4200
0.00130	3.0	2050	3550
0.00170	2.3	2200	3200
0.00215	1.8	2230	-
0.00240	-	2240	2800

The characteristics are shown in Figure 2. The initial positions of the characteristics are chosen so that the region of interest is covered. Since the velocity characteristics

have a steeper slope as shown in Figure 2, it is necessary to use characteristics originating both in the original propellant chamber and also at the projectile base during the motion.

The velocity boundary layer thickness, δ, is shown in Figure 3. The boundary layer grows not only with time but also with increase in x except near the base of the projectile. The growth is steeper at the projectile base than near the breech end. This analysis predicts the growth and also the decay continuously without any separate treatment. The boundary layers are much thicker than predicted by Nordheim, Soodak, and Nordheim[7]. It is straightforward to compute any other parameter of the boundary layer from this velocity boundary layer thickness parameter.

The dimensionless skin friction coefficient is shown in Figure 4. This is maximum near the breech end and decreases in the flow direction. The skin friction coefficient decreases also with increase in time. This also implies that the unsteady flows are more susceptible to viscous effects than steady flows. Of course, the shear stress at the wall can be computed by the use of skin friction coefficient and equation (9).

The distribution of shear stresses at the wall in PSI units are shown in Figure 5. The trend is somewhat opposite to that of the skin friction coefficient, i.e., the shear stresses are maximum near the projectile base. The shear stresses first increase and then decrease with increase in time. The shear stress at the wall will be maximum at the muzzle end of the gun barrel. Therefore, the gas dynamic erosion will be more severe at the muzzle end than near the origin of rifling. Even though the shear stresses are of the order of 100 psi and may not create significant erosion on high strength material such as steel, these shear stresses can easily wipe-off any incipient melting of the bore surface material.

ACKNOWLEDGMENT

The author would like to express appreciation to Miss Gigi Anderson for her skillful typing of the manuscript.

REFERENCES

1. Piquet, J., and R.K. Zeytounian, "Recent Research in the Field of Unsteady Boundary Layers," Proceedings of the International Union of Theoretical and Applied Mechanics, Laval University, Quebec, Canada, May 1971.

2. Reddy, K.C., W.L. Sickles, and R. Yalamanchili, "Computation of Unsteady Boundary Layers," Proc. of Unsteady Aerodynamics Symposium, University of Arizona, Tucson, Arizona, Mar. 1975.

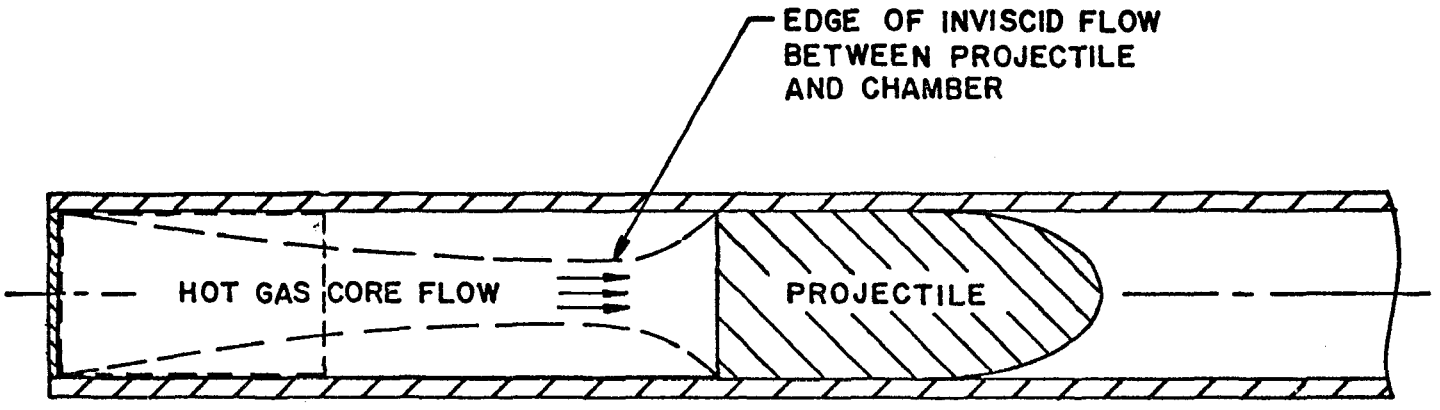
3. Yalamanchili, R., and P.D. Benzkofer, "Unsteady Compressible Boundary Layers With Arbitrary Pressure Gradients," AIAA Paper #73-132, 11th Aerospace Sciences Meeting, Washington, D.C., Jan. 1973 (also AD746235).

4. Yalamanchili, R., "Shock-Induced Boundary Layers By Weighted-Residuals and Method of Lines," Proc. of First International Conference on Computational Methods in Nonlinear Mechanics, TICOM, Univ. of Texas-Austin, Sept. 1974.

5. Yalamanchili, R., and A.J. Erickson, "Unsteady Turbulent Boundary Layers With Arbitrary Pressure Gradients," ASME Paper #74-WA/HT-34, ASME Winter Annual Meeting, New York City, N.Y., 1974.

6. Nikuradse, J., "Boundary Layer Theory," by H. Schlichting, McGraw-Hill, 6th Edition, 1968.

7. Nordheim, L.W., Soodak, H., and Nordheim, G., "Thermal Effects of Propellant Gases in Erosion Vents and In Guns," National Defense Research Committee Armor and Ordnance Report #A-262(SRD #3447), Division 1, May 1944.



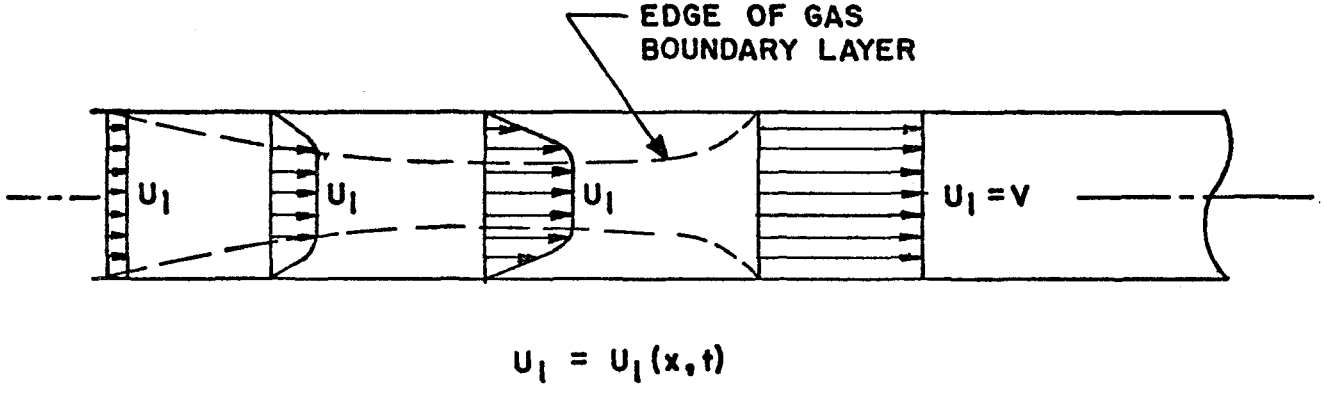


FIGURE 1 SCHEMATIC OF GUN BARREL AND ASSOCIATED VELOCITY PROFILES

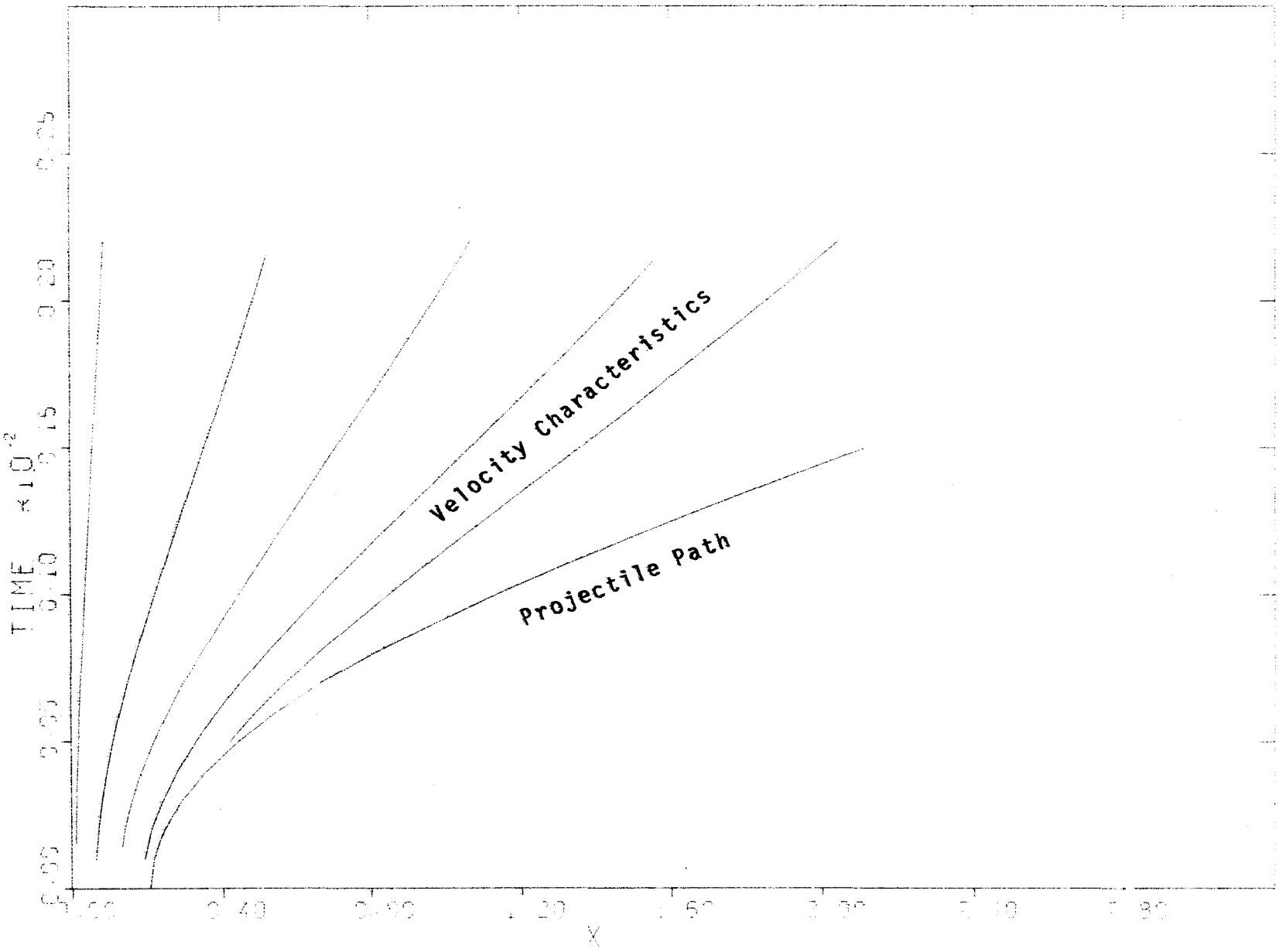


FIG. 2 PATH OF VELOCITY CHARACTERISTICS AND PROJECTILE

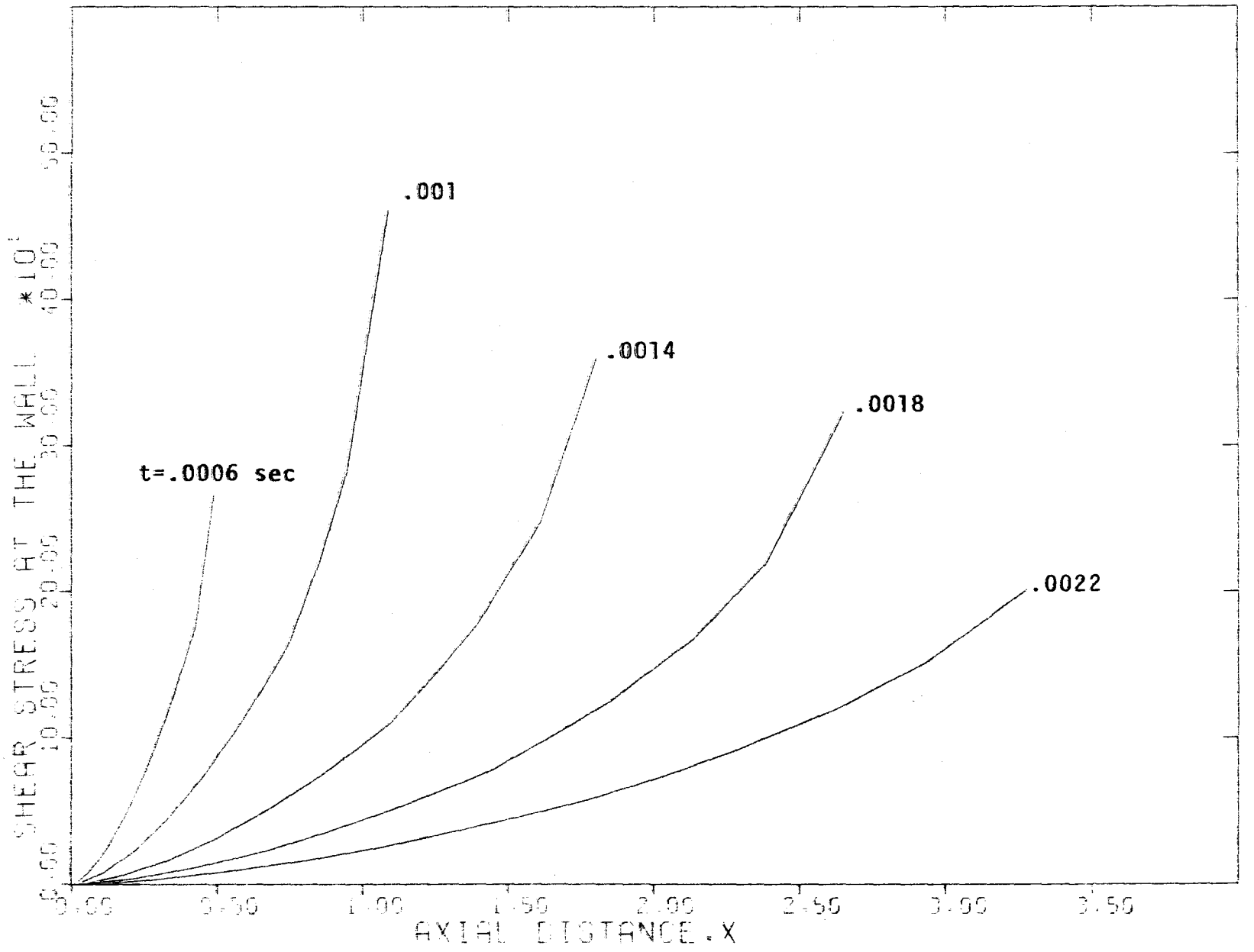


FIG. 5 DISTRIBUTION OF SHEAR STRESS OF THE WALL

NONLINEAR PROBLEMS IN THE INTERACTION OF A STEEL CARTRIDGE CASE AND THE CHAMBER

J. J. Toal and S. C. Chu

Research Directorate
GEN Thomas J. Rodman Laboratory
Rock Island Arsenal
Rock Island, Illinois 61201

ABSTRACT

In a recent development of a steel cartridge case, high extraction forces were encountered that were caused by the sticking of the case to the chamber. This sticking condition arose because a steel cartridge case has less recovery than a brass case. This reduced recovery can be attributed to the modulus of elasticity of steel that is much higher than the modulus of elasticity of brass. A relatively simple nonlinear elastoplastic analysis has been developed to parametrically analyze the cartridge case/chamber interaction under actual firing conditions when the cartridge case is loaded near the maximum material-carrying capacity. In contrast to the usual assumption that the chamber is rigid, the chamber is considered to be deformable in this investigation. The interaction of the steel cartridge and the chamber is studied parametrically. With the use of the analysis proposed in this paper, a set of design parameters can be found to ensure uniform and acceptable performance of a steel cartridge case.

INTRODUCTION

Cartridge cases for small-arms ammunition have been traditionally designed for brass. However, because of the limited natural supply of brass and its predominant use in small arms ammunition, a strategic significance has been attached to it. With the increasing emphasis of firepower and inadequate domestic supply, the usage of brass during a major war could become critical. Hence, the use of alternate materials for cartridge cases of small-arms ammunition is very important. Aluminum and steel are considered to be more economical materials than brass for this application. The major difficulties in the development of aluminum cartridge cases are those of the so-called "burn-through" problems. Existing literature indicates that the sidewall of an aluminum cartridge case that split during firing caused serious erosion of the case head and the chamber of a gun. Serious erosion damage in the M16 Rifle chamber due to aluminum case splits has been observed [Ref 1]. Hence, from the economical point of view, the development of steel-cased cartridges is of great importance at the present time. In the recent development of steel cartridge cases, the problem has arisen in which the cartridge case becomes stuck in the chamber, resulting in high extraction loads. The main cause of the "sticking" problem is

that a steel cartridge case has less recovery than a brass case since the modulus of elasticity of steel is much higher than the modulus of elasticity of brass.

The objective of this investigation was to develop a relatively simple nonlinear elastoplastic method for analyzing and designing a steel-cased cartridge to remedy the sticking problem. The interaction of a steel-cased cartridge and the chamber of a gun will be studied parametrically. A set of design parameters such as cartridge case material properties, chamber pressure, cartridge configuration, the initial clearance between a case and a chamber, and the configuration of the chamber of a gun, will be established to ensure uniform and acceptable performance of a steel-cased cartridge. In contrast to the usual assumption that the chamber is rigid, the chamber is considered to be deformable in this investigation.

THEORY

An inelastic theory for a cartridge case is developed on the basis of the following assumptions:

a. The material is homogeneous and isotropic.

b. The inelastic deformation is time-independent.

c. Hencky's stress-strain relation is valid.

d. The material is assumed to be compressible.

e. Von Mises' yield criterion is valid.

The use of cylindrical coordinates (r,θ,z) is convenient where z is coincident with the longitudinal axis of the case. For any point in a cartridge, the nonvanishing stress components are σ_r, σ_θ, and σ_z while the nonvanishing strain components are ε_r, ε_θ, and ε_z.

Elastoplastic Solution of Case-Chamber Problem

A longitudinal cross section of a cartridge case is shown in Figure 1. The cartridge case is divided into N rings with $z_0=\ell$, z_1, $z_2,\cdots z_n=L$. Let F_o be the bullet-pulling force. The load acting on the base of a bullet is given by

$$F = \pi r_2^2 P_g \tag{1}$$

where P_g is propellant gas pressure.

The stress components acting on a ring with radius r are

$$\sigma_\theta = \frac{Pr}{t\cos\beta} \tag{2}$$

$$\sigma_r = \frac{t}{2r}\ \sigma_\theta = -\frac{P}{2\cos\beta} \tag{3}$$

$$\sigma_z = \frac{Pr}{2t\cos\beta} \qquad \text{if } F \leq F_o \tag{4}$$

$$= \frac{P(r_1^2-r_2^2)}{2rt\cos\beta} \qquad \text{if } F > F_o \tag{5}$$

Note that $P = P_g - P_I$ where P_I is the interface pressure between a cartridge case and the chamber.

Cartridge Case Loaded Into Inelastic Region:

In the development of an elastoplastic solution, a loading function must be specified in which the plastic deformation and the subsequential yield condition are defined. For linear strain-hardening materials, the loading fucntion of Hill [Ref 2] can be written:

$$\bar{\sigma} = (1-\alpha)\sigma_{yc} + \alpha E_c \bar{\varepsilon} \tag{6}$$

where αE_c is the slope of the straight line in the plastic region and α may be considered as a strain-hardening factor for the material, σ_{yc} is the yield stress of the cartridge case material in tension, $\bar{\sigma}$ and $\bar{\varepsilon}$ are effective stress and effective strain, respectively, which are defined as

$$\bar{\sigma} = \frac{1}{\sqrt{2}} [(\sigma_r-\sigma_\theta)^2 + (\sigma_\theta-\sigma_z)^2 + (\sigma_z-\sigma_r)^2]^{\frac{1}{2}} \tag{7}$$

and

$$\bar{\varepsilon} = \frac{\sqrt{2}}{3} [(\varepsilon_r-\varepsilon_\theta)^2 + (\varepsilon_\theta-\varepsilon_z)^2 + (\varepsilon_z-\varepsilon_r)^2]^{\frac{1}{2}} \tag{8}$$

since $\bar{\varepsilon} = \bar{\varepsilon}^e + \bar{\varepsilon}^p$, equation (6) can be reduced to the relation [Ref 3]

$$\bar{\sigma} = \sigma_{yc} + \frac{\alpha E_c}{1-\sigma} \bar{\varepsilon}^p \tag{9}$$

or

$$\bar{\varepsilon}^p = \frac{1-\alpha}{\alpha E_c} (\bar{\sigma} - \sigma_{yc}) \tag{10}$$

With the use of Hencky's stress-strain relation [Ref 4],

$$\varepsilon_{ij}^p = \frac{3}{2} \frac{S_{ij}}{\bar{\sigma}} \bar{\varepsilon}^p \tag{11}$$

where S_{ij} is a deviatoric stress tensor. Hence,

$$\varepsilon_\theta^p = \frac{3}{2}\frac{\bar{\varepsilon}^p}{\bar{\sigma}} S_\theta = \frac{1}{2}\frac{\bar{\varepsilon}^p}{\bar{\sigma}}[2\sigma_\theta-\sigma_r-\sigma_z] \tag{12}$$

Substituting eq. (10) into eq. (12) yields

$$\varepsilon_\theta^p = \frac{(1-\alpha)}{2\alpha E_c}\left(1-\frac{\sigma_{yc}}{\bar{\sigma}}\right)(2\sigma_\theta-\sigma_r-\sigma_z) \tag{13}$$

With the use of the strain-deflection relation

$$\varepsilon_\theta = \frac{u_c}{r} \tag{14}$$

where u_c is the radial displacement of the cartridge case at radius r, ε_θ is the circumferential strain which is the sum of the elastic part and the inelastic part, i.e.,

$$\varepsilon_\theta = \varepsilon_\theta^e + \varepsilon_\theta^p \tag{15}$$

Hence, equation (14) can be written

$$u_c = r(\varepsilon_\theta^e + \varepsilon_\theta^p) \tag{16}$$

From the theory of elasticity, one obtains the following equation

$$\varepsilon_\theta^e = \frac{1}{E_c}[\sigma_\theta - \nu_c(\sigma_r + \sigma_z)] \tag{17}$$

Substituting eqs. (13) and (17) into eq. (16), then

$$u_c = r\{\frac{1}{E_c}[\sigma_\theta-\nu_c(\sigma_r+\sigma_z)] + \frac{(1-\alpha)}{2\alpha E_c}\left(1-\frac{\sigma_{yc}}{\bar{\sigma}}\right)(2\sigma_\theta-\sigma_r-\sigma_z)\} \tag{18}$$

Let e be the initial clearance between cartridge case and chamber. Then the displacement of the chamber at the inner surface, u_b, can be written

$$u_b = u_c - e \tag{19}$$

If the chamber is assumed to be a thick-walled cylinder subjected to internal pressure P_I, then one has [Ref 5]

$$u_b = u_c - e = \frac{P_I a}{E_b}\left[\frac{b^2+a^2}{b^2-a^2} + \nu_b\right] \tag{20}$$

Hence,

$$P_I = \frac{(u_c - e)\, E_b}{a\left[\frac{b^2+a^2}{b^2-a^2} + \nu_b\right]} \qquad \text{if } u_c > e \tag{21}$$

$$P_I = 0 \qquad \text{if } u_c \leq e \tag{22}$$

Note that:

$$P_g = P + P_I \qquad \text{if } u_c > e \tag{23}$$

$$P_g = P \qquad \text{if } u_c \leq e \tag{24}$$

Cartridge Case Loaded In Elastic Region:

If the cartridge case is loaded in the elastic region, i.e., $\bar{\sigma} < \sigma_{yc}$, then the following stress-strain and strain-displacement relations will be used

$$\begin{aligned} \varepsilon_r &= \frac{1}{E}\left[\sigma_r - \nu_c\,(\sigma_\theta + \sigma_z)\right] \\ \varepsilon_\theta &= \frac{1}{E_c}\left[\sigma_\theta - \nu_c\,(\sigma_r + \sigma_z)\right] \\ \varepsilon_z &= \frac{1}{E_c}\left[\sigma_z - \nu_c\,(\sigma_r + \sigma_\theta)\right] \end{aligned} \tag{25}$$

and

$$u_c = r\varepsilon_\theta = \frac{r}{E_c}\left[\sigma_\theta - \nu_c\,(\sigma_r + \sigma_z)\right] \tag{26}$$

During the unloading process, the behavior of the cartridge case and the chamber is considered to be elastic. The dimensions used are the dimensions of the case and chamber computed at the peak pressure. A negative pressure is applied to the case during the unloading process. The result of unloading is then superposed on the result at the peak pressure and, hence, the interference pressure and the extraction force can be determined.

COMPUTATIONAL METHOD

Loading

Since a large displacement of the cartridge case may be involved in the present case-chamber interface problem, the true displacement of the cartridge case at the peak pressure, P_{max}, would not be known at the beginning of the loading process. Hence, an incremental loading procedure will be utilized to obtain solutions in this investigation.

For each increment of pressure, the calculation procedure can be stated as follows:

Step 1 Specify an increment of pressure ΔP_i acting on the case, then:

$$P_i = P_{i-1} + \Delta P_i$$

Step 2 Calculate σ_θ, σ_r, and σ_z by using eqs. (2), (3), and (4) or (5).

Step 3 Calculate effective stress, $\bar{\sigma}$, by using eq. (7). Note that

if $\bar{\sigma} < \sigma_{yc}$, case loaded in elastic region

if $\sigma \geq \sigma_{yc}$, case loaded into inelastic region

Step 4 If $\bar{\sigma} < \sigma_{yc}$, calculate circumferential strain

$$\varepsilon_\theta = \frac{1}{E_c} [\sigma_\theta - \nu_c(\sigma_r + \sigma_z)]$$

Step 5 If $\bar{\sigma} \geq \sigma_{yc}$, calculate effective plastic strain

$$\bar{\varepsilon}^p = \frac{1-\alpha}{\alpha E_c} (\bar{\sigma} - \sigma_{yc})$$

Step 6 Calculate plastic component of circumferential strain

$$\varepsilon_\theta^p = \frac{1}{2} \frac{\bar{\varepsilon}^p}{\bar{\sigma}} [2\sigma_\theta - \sigma_r - \sigma_z]$$

Step 7 Compute radial displacement of the case

$$u_c = r\{\frac{1}{E_c} [\sigma_\theta - \nu_c(\sigma_r + \sigma_z)] + \varepsilon_\theta^p\}$$

Step 8 Compute bore displacement of the chamber

$u_b = u_c - e$ if $u_c > e$

$= 0$ if $u_c \leq e$

where e is the initial clearance between case and chamber

Step 9 Compute interface pressure

$$P_I = \frac{(u_c - e)E_b}{a\left[\frac{b^2+a^2}{b^2-a^2} + \nu_b\right]}$$

Step 10 Compute propellant gas pressure

$P_g = P_i + P_I$

If $P_g < P_{max}$, let $r = r_o + u_c$, then refer back to Step 1. All computations to continue.

Step 11 If $P_g = P_{max}$, then let

$\bar{P}_I = P_I$

$\bar{u}_c = u_c$

$\bar{u}_b = u_b$

$r_{max} = r_o + \bar{u}_c$

Hence, $\bar{P}_I$, $\bar{u}_c$, $\bar{u}_b$, and r_{max} are interface pressure between the case and the chamber, displacement of the case, displacement of the chamber at bore, and radius of the case at the peak pressure, respectively.

Unloading

Step 1 Specify $-\Delta P_i$, then calculate P_i by

$$P_i = P_{i-1} - \Delta P_i$$

Step 2 Compute σ_r, σ_θ, and σ_z

Step 3 Compute

$$\varepsilon_\theta = \frac{1}{E_c}\left[\sigma_\theta - \nu_c(\sigma_r + \sigma_z)\right]$$

Step 4 Compute case radial displacement

$$u_c = \frac{r}{E}\left[\sigma_\theta - \nu_c(\sigma_r + \sigma_z)\right]$$

Step 5 Compute interface pressure

$$P_I = \frac{u_c E_b}{a\left[\frac{b^2+a^2}{b^2-a^2} + \nu_b\right]}$$

Step 6 Applying the superposition principle, the interface pressure between case and chamber and the case and chamber displacements at unloading can be computed by the relations of

$$\bar{\bar{P}} = P_I + \bar{P}_I \qquad \text{if } |u_c| < \bar{u}_b$$

$$= 0 \qquad \text{if } |u_c| \geq \bar{u}_b$$

$$\bar{\bar{u}}_c = u_c + \bar{u}_c$$

$$\bar{\bar{u}}_b = u_b + \bar{u}_b \qquad \text{if } |u_c| \leq \bar{u}_b$$

$$= 0 \qquad \text{if } |u_c| > \bar{u}_b$$

where $\bar{\bar{P}}$ is the interface pressure during unloading

$\bar{\bar{u}}$ is the case displacement during unloading

and $\bar{\bar{u}}_b$ is the bore displacement of chamber during unloading

Step 7 Compute

$P_g = P_i + P_I$

if $P_g + P_{max} > 0$, then let $r = r_{max} + u_c$, refer back to Step 1. All computations to continue.

Step 8 If $P_g + \bar{P}_g \approx 0$, then compute extraction force. Note that when the unloading process is completed, the interface pressure is automatically computed at Step 6 which is used to compute extraction force with assigned friction coefficient between the case and the chamber.

ANALYSIS OF RESULTS

The steel cartridge case considered in this investigation is shown in Figure 1. The outer diameter of the chamber is assumed to be one inch. To obtain numerical solutions, the cartridge case is divided into 32 rings. The thickness; the initial clearance between case and chamber; material properties such as yield strength, modulus of elasticity, strain-hardening factors, etc., for each ring; and the peak pressure are considered to be independent design parameters. The effects of those parameters on the extraction forces are analyzed. A steel cartridge case is considered in this investigation, and since the modulus of elasticity of steel does not vary much, it will be considered as a constant, $E = 30 \times 10^6$ psi, for each ring.

Effect of the Peak Pressure

The magnitude of the peak pressure is a very important factor in causing the sticking problem of the steel cartridge case. The effect of the peak pressure on the resulting extraction force is shown in Figure 2. In this figure, note that the incidence of cases that stick to the chamber can be significantly reduced or eliminated if the peak pressure can be reduced to a certain level. A reduction of peak pressure (all other factors held constant) will reduce the round impulse and velocity. One method of maintaining the round impulse and the velocity while the peak pressure is being reduced is to cause the Pressure-Time (P-T) curve to become flatter.

Effect of Yield Strength of Case Material

The material properties such as yield strength of a steel-cased cartridge also play a very important role in the sticking problem. The effect of yield strength of a case material on the extraction force is shown in Figure 3. The extraction force can be reduced if a case is made of a higher strength material, as indicated in Figure 3. However, the increasing of the strength of material is limited by other material properties such as hardness and elongation. If the hardness of a cartridge case is increased, then the ductility of material will be reduced and hence the incidence of a split case will be increased.

Effect of Initial Clearance Between Cartridge Case and Chamber

The relation of clearance between cartridge case and chamber, and the extraction force is shown graphically in Figure 4. Note that the increasing of an initial clearance will reduce the force required to extract a case that is sticking to the chamber. However, this increasing of the initial clearance may increase the incidence of malfunctions in feeding.

CONCLUSIONS

A simple nonlinear solution technique has been presented for solving problems in the interaction of a steel cartridge case and a chamber. Both nonlinear material response and geometric nonlinearity have been taken into consideration in this investigation. Material nonlinearity has been taken into account by use of the theory of plasticity and Hencky's constitutive equations. An incremental loading procedure has been used in consideration of large deformation of a cartridge case.

Guidance for the selection of design variables such as yield strength of cartridge case material, peak chamber pressure, and initial clearance between cartridge case and chamber is presented graphically in this paper.

REFERENCES

1. Skochko, L., Rosenbaum, M., and Donnard, R., "Aluminum Cartridge Case, Concepts Task-Work Summary," Report R-3001, Frankford Arsenal, March 1974.

2. Hill, R., The Mathematical Theory of Plasticity, Oxford at the Clarendon Press, 1950.

3. Smith, J.O., and Sidebottom, O.M., Inelastic Behavior of Load-Carrying Members, John Wiley and Sons, Inc., 1965.

4. Mendelson, A., Plasticity: Theory and Application, The MacMillan Company, 1968.

5. Timoshenko, S., and Goodier, J.N., Theory of Elasticity, McGraw-Hill, New York, 1951.

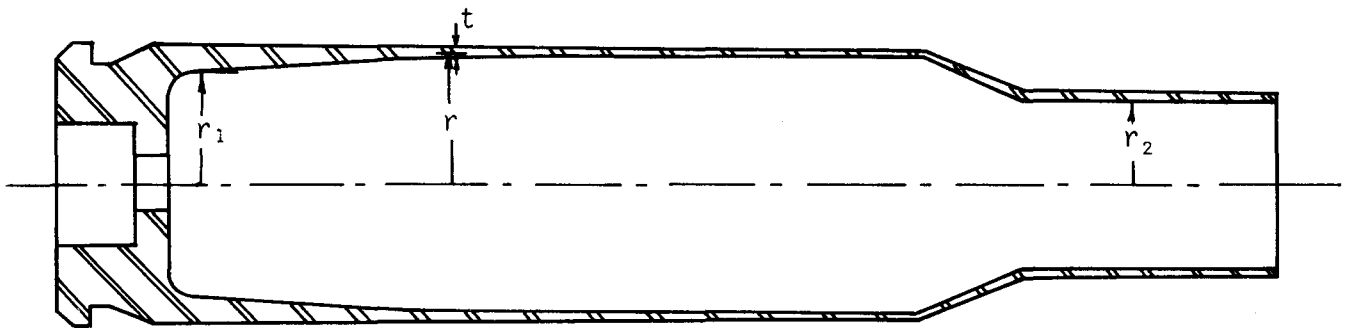


Fig. 1 Longitudinal Cross Section of A Cartridge Case

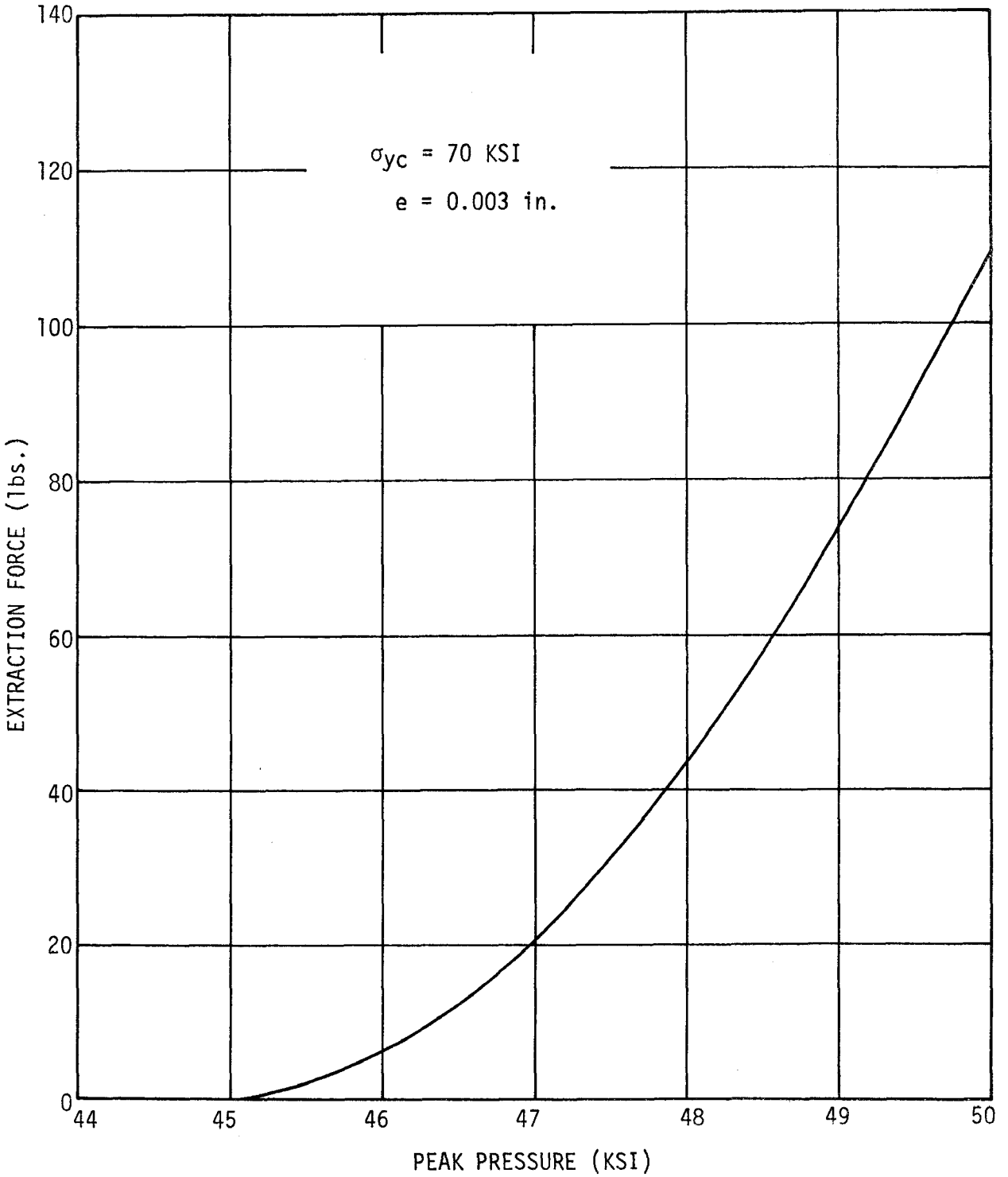


Fig. 2 Effect of the Peak Pressure On The Extraction Force

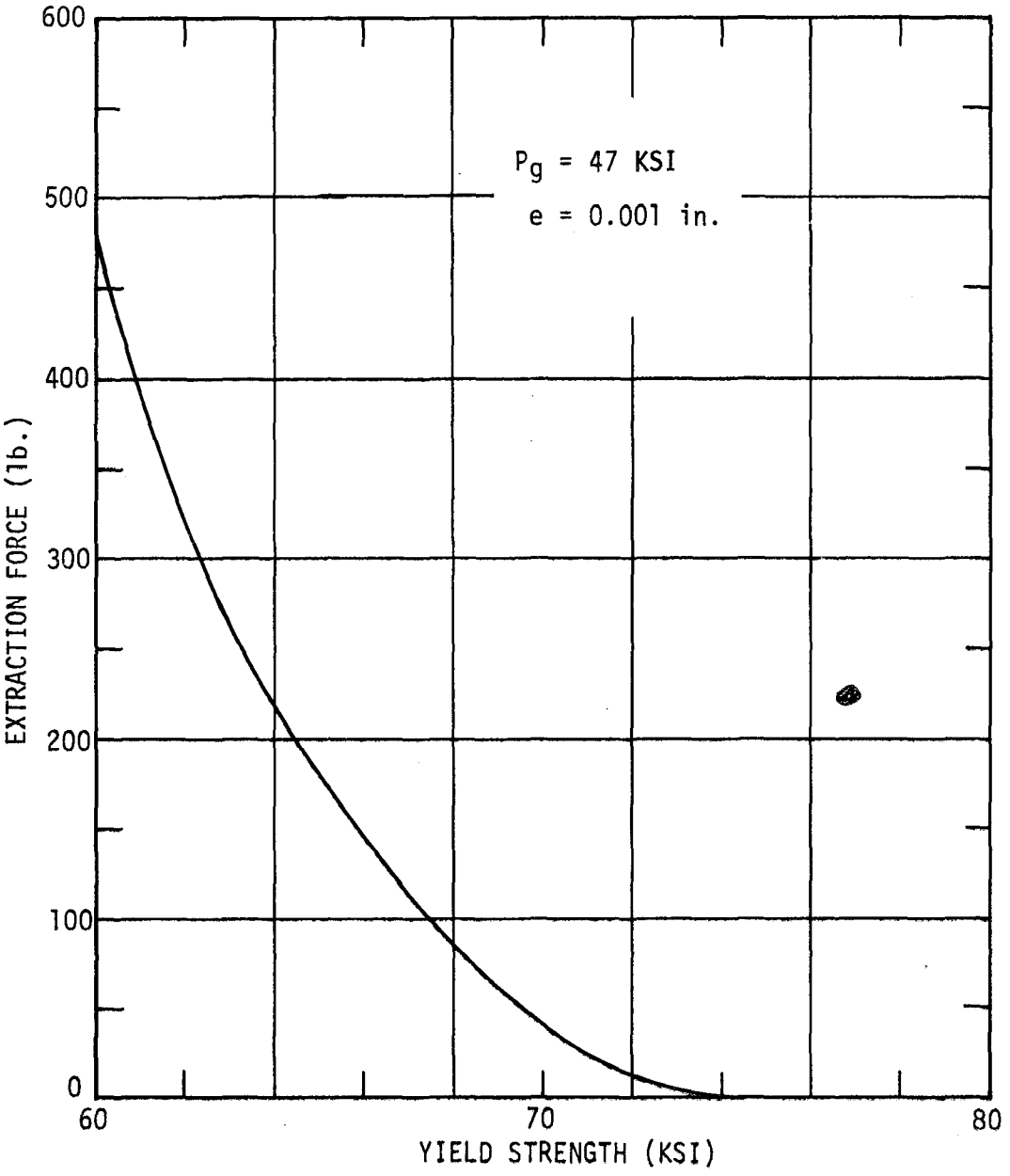


Fig. 3 Effect of the Yield Strength of the Case Material On The Extraction Force

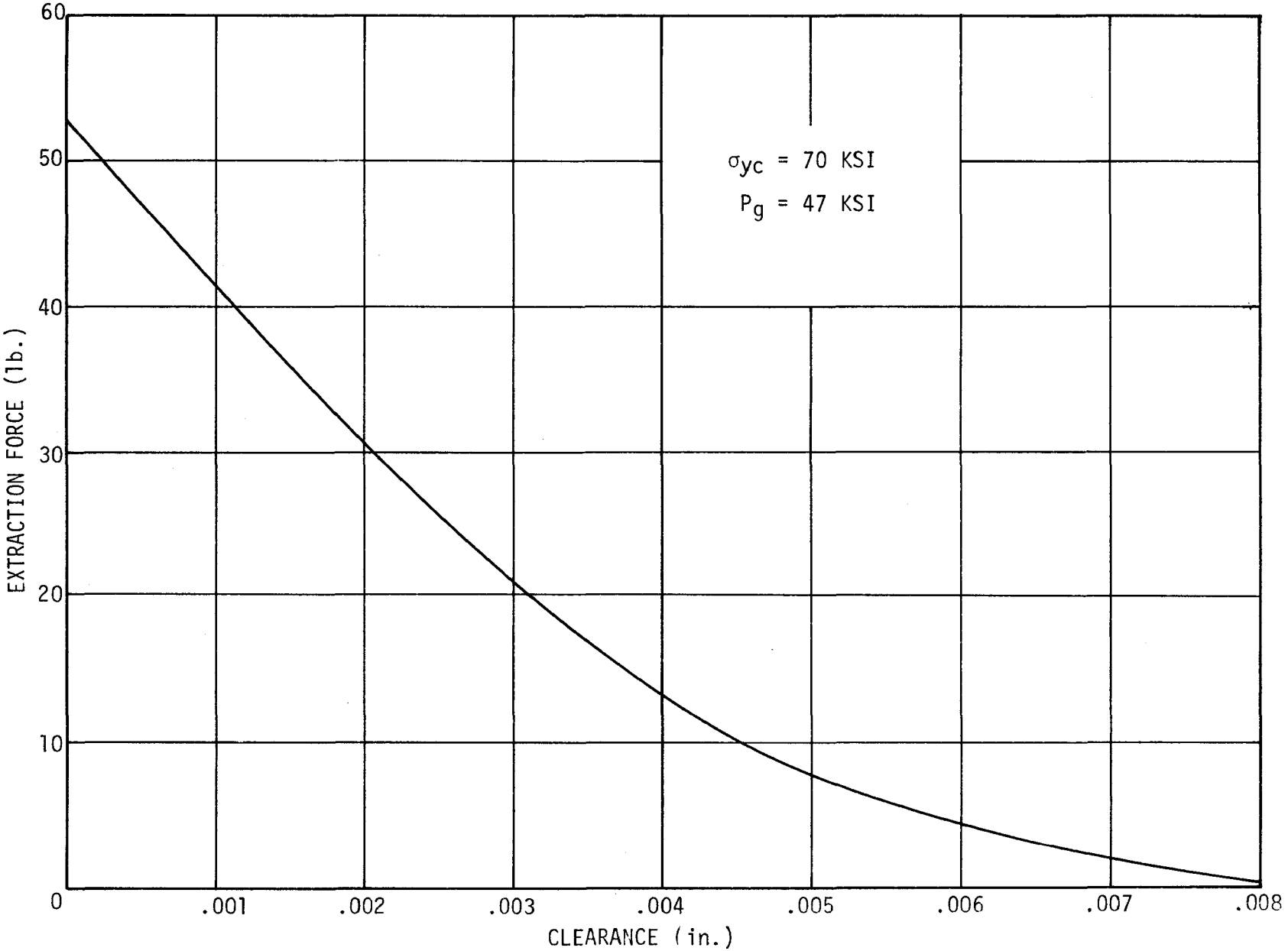


Fig. 4 Effect of the Clearance Between Case and Chamber On The Extraction Force

NINE-POINT DIFFERENCE SOLUTIONS FOR POISSON'S EQUATION

J. Barkley Rosser
Professor of Mathematics
Mathematics Research Center
University of Wisconsin-Madison
Madison, Wisconsin 53706

Abstract. It is shown that "stencils" exist for the sixth order solution of Poisson's equation by use of a nine-point difference approximation. This enables one to get more accurate approximations for the solution with less labor.

NINE-POINT DIFFERENCE SOLUTIONS FOR POISSON'S EQUATION

J. Barkley Rosser

Introduction. Suppose we wish to approximate the solution of

$$\nabla^2 u(x, y) = f(x, y) \tag{1.1}$$

inside some region, given values around the boundary. A standard approach is to introduce a "stencil", say

0	1	0
1	-4	1
0	1	0

$u(x, y)$.

The significance of this is that if a coefficient in the stencil is m units above the horizontal center line and n units to the right of the vertical center line (m and/or n may be negative), one forms the product of the coefficient with $u(x + nh, y + mk)$; the entire stencil denotes the sum of these products. Thus the stencil shown above denotes

$$u(x + h, y) + u(x - h, y) + u(x, y + k) + u(x, y - k) - 4u(x, y) .$$

To approximate the $u(x, y)$ which solves (1.1) it is traditional to choose $k = h$. Also, to save space, we write

Sponsored by the United States Army under Contract No. DA-31-124-ARO-D-462.

(1.2) $$\Delta_5 u(x, y) = \begin{array}{|c|c|c|} \hline 0 & 1 & 0 \\ \hline 1 & -4 & 1 \\ \hline 0 & 1 & 0 \\ \hline \end{array}\ u(x, y) .$$

This is called the five-point stencil.

If $u(x, y)$ is reasonably smooth, we have

(1.3) $$\Delta_5 u(x, y) \cong h^2 \nabla^2 u(x, y)$$

to order h^4. So, if h is reasonably small,

(1.4) $$\Delta_5 u(x, y) \cong h^2 f(x, y)$$

is a good approximation to (1.1). Let us choose a suitable (x_0, y_0) and define

(1.5) $$u_{m,n} = u(x_0 + mh, y_0 + nh) .$$

Then to approximation h^4, we conclude

(1.6) $$\Delta_5 u_{m,n} \cong h^2 f(x_0 + mh, y_0 + nh) .$$

The equations (1.6) are a set of linear equations, which have a unique exact solution, $\bar{u}_{m,n}$. If we solve these linear equations, we will get quantities $\bar{u}_{m,n}$ such that $\bar{u}_{m,n}$ differs from $u_{m,n}$ by order h^2 (assuming smooth boundary conditions).

This is the basis for several schemes for computing numerical approximations for $u(x, y)$ at the "grid points" $(x_0 + mh, y_0 + nh)$. However, since the accuracy is only to order h^2, it is not possible to get very high accuracy. Even moderate accuracy requires solving a very large number of simultaneous equations.

If $f(x, y) = 0$, one can get higher order approximations by use of a nine-point stencil, and so improve the situation. The nine-point stencil is given by

(1.7) $$\Delta_9 u(x, y) = \begin{array}{|c|c|c|} \hline 1 & 4 & 1 \\ \hline 4 & -20 & 4 \\ \hline 1 & 4 & 1 \\ \hline \end{array}\, u(x, y) .$$

In general, one has

(1.8) $$\Delta_9 u(x, y) \cong 6h^2 \nabla^2 u(x, y)$$

only to order h^4. So if one solves

(1.9) $$\Delta_9 \bar{u}_{m,n} = 6h^2 f(x_0 + mh, y_0 + nh) ,$$

one will still usually only get an approximation to order h^2. However, if one solves

(1.10) $$\Delta_9 \bar{u}_{m,n} = 0 ,$$

one will get an approximation to order h^6 for the solution of

$$\nabla^2 u(x, y) = 0 , \tag{1.11}$$

PROVIDED one has smooth enough boundary conditions.

The object of the present paper is to present methods for solving (1.1) by a nine-point stencil to order h^6 even when $f(x, y)$ is not identically zero. We should warn that these methods will fail unless the boundary conditions and $f(x, y)$ are smooth. In particular, $f(x, y)$ should have bounded derivatives up to order six for the methods of this paper to succeed.

For completeness, we repeat certain material from Rosser [1].

2. A fourth order method. In Section 1, we contemplated dividing our region into squares. For some types of regions, it would be convenient to divide the region into rectangles. It is widely believed that difference methods cannot be constructed to give approximations of order greater than two unless the region is divided into squares. This is not so. We will explain a method that gives approximations of order four if rectangles of sides h and k are used.

Let us temporarily set

$$u_{m, n} = u(x_0 + mh, y_0 + nk) . \tag{2.1}$$

That is, we use rectangles whose corners are the grid points $(x_0 + mh, y_0 + nk)$.

Define

$$A = \frac{12h^2k^2}{h^2 + k^2} \tag{2.2}$$

$$b = \frac{10k^2 - 2h^2}{h^2 + k^2} \tag{2.3}$$

$$c = \frac{10h^2 - 2k^2}{h^2 + k^2} \; . \tag{2.4}$$

We have

$$(b + 2)h^2 = A = (c + 2)k^2 \; , \tag{2.5}$$

$$b + c = 8 \; . \tag{2.6}$$

Define a modified nine-point stencil

$$\Delta_9^* u(x, y) = \begin{array}{|c|c|c|} \hline 1 & c & 1 \\ \hline b & -20 & b \\ \hline 1 & c & 1 \\ \hline \end{array} \; u(x, y) \; ; \tag{2.7}$$

here motion of one unit in the y-direction in the stencil is supposed to induce a change of k in y, as in our original definition.

If $u(x, y)$ is smooth, we have to order $h^6 + k^6$

$$\Delta_9^* u(x, y) \cong A\nabla^2 u(x, y) + \frac{h^2 A}{12} u_{xxxx}(x, y) + h^2k^2 u_{xxyy}(x, y) + \frac{k^2 A}{12} u_{yyyy}(x, y) \; . \tag{2.8}$$

By (1.1) we have

$$u_{xxxx}(x, y) + u_{xxyy}(x, y) = f_{xx}(x, y)$$

$$u_{xxyy}(x, y) + u_{yyyy}(x, y) = f_{yy}(x, y) .$$

If we multiply the first of these by $h^2A/12$ and the second by $k^2A/12$ and add, we see by (2.2) that we can write (2.8) as

$$\Delta_9^* u(x, y) \cong Af(x, y) + \frac{h^2A}{12} f_{xx}(x, y) + \frac{k^2A}{12} f_{yy}(x, y) . \tag{2.9}$$

Observe that

$$h^2 f_{xx}(x, y) \cong f(x + h, y) + f(x - h, y) - 2f(x, y)$$

$$k^2 f_{yy}(x, y) \cong f(x, y + k) + f(x, y - k) - 2f(x, y) .$$

Thus we conclude finally that to order $h^6 + k^6$

$$\Delta_9^* u(x, y) \cong \frac{h^2k^2}{h^2 + k^2} \begin{array}{|c|c|c|} \hline 0 & 1 & 0 \\ \hline 1 & 8 & 1 \\ \hline 0 & 1 & 0 \\ \hline \end{array} f(x, y) . \tag{2.10}$$

So, if we solve

$$(2.11)\qquad \Delta_9^{*-}u_{m,n} = \frac{h^2k^2}{h^2+k^2}\begin{array}{|c|c|c|}\hline 0 & 1 & 0 \\ \hline 1 & 8 & 1 \\ \hline 0 & 1 & 0 \\ \hline\end{array} f(x_0 + mh, y_0 + nk)\ ,$$

we will get $\bar{u}_{m,n}$ that differ from $u_{m,n}$ by the order $h^4 + k^4$.

This gives a method of order four for rectangular grid elements.

3. A sixth order method. If there is a sixth order method that permits the use of rectangular grid elements, we have no knowledge of it. So we return to square grid elements, adopting again the notations of Section 1.

If (1.1) holds and $u(x,y)$ is smooth enough, then to order h^8

$$(3.1)\qquad \Delta_9 u(x,y) \cong 6h^2 f(x,y) + \frac{h^4}{2}\nabla^2 f(x,y) + \frac{h^6}{60}\nabla^4 f(x,y) + \frac{h^6}{30} f_{xxyy}(x,y)\ .$$

This does not agree exactly with equation (20.57) on p. 194 of Forsythe and Wasow [2]. However, they claim that their (20.57) is copied from another reference, but they made a mistake in copying. Our (3.1) agrees with the formula from which (20.57) was supposed to be copied.

We note that

(3.2)

1	1	1
1	-8	1
1	1	1

$$f(x, y) \cong 3h^2\nabla^2 f(x, y) + \frac{h^4}{4}\nabla^4 f(x, y) + \frac{h^4}{2} f_{xxyy}(x, y)$$

to order h^6, if $f(x, y)$ is sufficiently smooth. So we may replace (3.1) by

(3.3)

1	1	1
1	82	1
1	1	1

$$\Delta_9 u(x, y) \cong \frac{h^2}{15} f(x, y) + \frac{3h^4}{10}\nabla^2 f(x, y) .$$

This formula appears in Collatz [3] as one of the stencils in Table VI on p. 543.

There are occasional circumstances in which this would be quite adequate; for instance if $f(x, y)$ is a harmonic function. However, in general we must do better.

For computational purposes, when we replace the $u_{m,n}$ by $\bar{u}_{m,n}$ and attempt to solve, it does not matter what is on the right side of our equations, as long as we maintain the same form on the left side. So there is no reason why we must restrict the right side of (3.3) to a nine-point stencil.

As we will have a solution to order h^6, we need not take h particularly small. Useful results have been obtained with $h = L/6$, where L is a basic dimension of the entire figure. With $h = L/10$ accurate results should result, and with $h = L/20$ quite high accuracy should result. For a square region, this requires at most 441 grid points, including the boundary. It would not overload the memory to compute $f(x, y)$ in advance at every grid point, store these values, and call such as are needed for each application of (3.3) or its replacement. In fact, it would be an efficient scheme of computation. So we use a larger stencil on the right of (3.3).

We note that

$$h^2 f_{xx}(x,y) \cong -\frac{1}{12} f(x+2h, y) + \frac{4}{3} f(x+h, y) - \frac{5}{2} f(x, y) + \frac{4}{3} f(x-h, y) - \frac{1}{12} f(x-2h, y) \tag{3.4}$$

to order h^6. By using this and the corresponding relation for $f_{yy}(x, y)$ in (3.3), we obtain

$$(3.5)\qquad \Delta_9 u(x,y) \cong \frac{h^2}{120}\begin{array}{|c|c|c|c|c|}\hline 0 & 0 & -3 & 0 & 0 \\ \hline 0 & 8 & 56 & 8 & 0 \\ \hline -3 & 56 & 476 & 56 & -3 \\ \hline 0 & 8 & 56 & 8 & 0 \\ \hline 0 & 0 & -3 & 0 & 0 \\ \hline \end{array} f(x,y)\ .$$

This will serve very satisfactorily except for points adjacent to the boundary. For these, values of $f(x, y)$ at points outside the region would be required. If such are available, there would be no difficulty. However, they might not be available.

There is of course the off-center difference approximation

$$(3.6)\qquad h^2 f_{xx}(x,y) \cong \frac{5}{6} f(x-h,y) - \frac{5}{4} f(x,y) - \frac{1}{3} f(x+h,y) + \frac{7}{6} f(x+2h,y)$$
$$- \frac{1}{2} f(x+3h,y) + \frac{1}{12} f(x+4h,y)\ ,$$

correct to order h^6. Using it, and the corresponding relation for y, we could derive

(3.7) $$\Delta_9 u(x,y) \cong \frac{h^2}{120}$$

0	3	0	0	0	0
0	-18	0	0	0	0
0	42	0	0	0	0
8	-4	8	0	0	0
38	566	-4	42	-18	3
8	38	8	0	0	0

$f(x,y)$,

which is valid to order h^8; on the right of (3.7) the "origin" of the stencil is the square which contains 566.

This certainly brings a method of order h^6 within reach. One wonders if there could be a better stencil than that on the right of (3.7). Perhaps there is not, but we will investigate what is available.

4. A general approach. In order to take care of grid points that are one unit away from each of two edges, it follows by (3.1) that we require constants $a_{m,n}$ such that

(4.1) $$\sum_{m=-1}^{S}\sum_{n=-1}^{S} a_{m,n} f(x+mh, y+nh) \cong 6f(x,y) + \frac{h^2}{2}\nabla^2 f(x,y)$$

$$+ \frac{h^4}{60}\nabla^4 f(x,y) + \frac{h^4}{30} f_{xxyy}(x,y)$$

to within terms of order h^6. It follows by (3.7) that such constants

exist for $S = 4$. We shall show that they do not exist for $S < 4$. For $S = 4$, there are many sets of $a_{m,n}$, and we shall derive the general form.

Because the right side of (4.1) is invariant under interchange of x and y, if $a_{m,n}$ satisfy (4.1), then so would $a^*_{m,n}$, where we take

$$a^*_{m,n} = a_{n,m} .$$

Then so would

$$a^{**}_{m,n} = \frac{1}{2}(a_{m,n} + a^*_{m,n}) .$$

So we lose no generality in assuming

$$a_{m,n} = a_{n,m} . \tag{4.2}$$

If $f(x, y)$ is smooth enough to have a double Taylor series out to order h^6, then

$$\sum_{m=-1}^{S} \sum_{n=-1}^{S} a_{m,n} f(x + mh, y + nh) \cong \sum_{r=0}^{5} h^r \sum_{s=0}^{r} \frac{K_{rs}}{s!(r-s)!} D_x^{r-s} D_y^s f(x, y) \tag{4.3}$$

to within terms of order h^6, where D_x and D_y are partial derivatives with respect to x and y respectively, and

$$K_{rs} = \sum_{m=-1}^{S} \sum_{n=-1}^{S} m^{r-s} n^s a_{m,n} . \tag{4.4}$$

Because of (4.2), we have

$$K_{rs} = K_{sr} . \tag{4.5}$$

Define

$$A_m = \sum_{n=-1}^{S} a_{m,n} . \tag{4.6}$$

By (4.4)

$$K_{r0} = \sum_{m=-1}^{S} m^r A_m \,. \tag{4.7}$$

By (4.3), if we are to satisfy (4.1), we must have

$$K_{00} = 6 \tag{4.8}$$

$$K_{10} = 0 \tag{4.9}$$

$$K_{20} = 1 \tag{4.10}$$

$$K_{30} = 0 \tag{4.11}$$

$$K_{40} = \frac{2}{5} \tag{4.12}$$

$$K_{50} = 0 \,. \tag{4.13}$$

By (4.7), this is a set of six simultaneous linear equations for the A_m. If $S < 4$, they have no solution. So we take $S = 4$, for which we observed earlier that there is a solution, and proceed. The equations (4.8) through (4.13) have the unique solution

$$A_{-1} = \frac{9}{20}, \; A_0 = \frac{209}{40}, \; A_1 = \frac{1}{10},$$

$$A_2 = \frac{7}{20}, \; A_3 = -\frac{3}{20}, \; A_4 = \frac{1}{40} \,.$$

Analogously, if we write

$$B_m = \sum_{n=-1}^{S} n a_{m,n}, \tag{4.14}$$

then to satisfy (4.1) we must have

$$\sum_{m=-1}^{S} m^{r-1} B_m = K_{r1} = 0 \qquad (1 \le r \le 5) \,. \tag{4.15}$$

This set of 5 equations has the one fold multiplicity of solutions

$$B_{-1} = -\frac{1}{5}B_0,\ B_1 = -2B_0,\ B_2 = 2B_0,\ B_3 = -B_0,\ B_4 = \frac{1}{5}B_0 .$$

We write also

$$C_m = \sum_{n=-1}^{S} n^2 a_{m,n} . \tag{4.16}$$

To satisfy (4.1) we must have

$$\sum_{m=-1}^{S} C_m = K_{22} = 1 \tag{4.17}$$

$$\sum_{m=-1}^{S} mC_m = K_{32} = 0 \tag{4.18}$$

$$\sum_{m=-1}^{S} m^2 C_m = K_{42} = \frac{4}{15} \tag{4.19}$$

$$\sum_{m=-1}^{S} m^3 C_m = K_{52} = 0 . \tag{4.20}$$

This set of 4 equations has the two fold multiplicity of solutions

$$C_{-1} = \frac{19}{60} - \frac{1}{4}C_0 + \frac{1}{4}C_4,\ C_1 = \frac{37}{30} - \frac{3}{2}C_0 - \frac{5}{2}C_4 ,$$

$$C_2 = -\frac{11}{15} + C_0 + 5C_4,\ C_3 = \frac{11}{60} - \frac{1}{4}C_0 - \frac{15}{4}C_4 .$$

To satisfy (4.1) it is sufficient as well as necessary to satisfy equations (4.8) - (4.13), (4.15), and (4.17) - (4.20). This we have accomplished, and with three parameters, B_0, C_0, and C_4, at our disposal.

Given values of the A_m, B_m, and C_m, we have yet to determine the $a_{m,n}$. By (4.2) and (4.6), we have

$$a_{m,0} = A_m - a_{m,-1} - \sum_{n=1}^{4} a_{m,n} . \tag{4.21}$$

Except for $m = 0$, this expresses $a_{m,0}$ (and hence $a_{0;m}$) in terms of $a_{r,s}$ with both $r \neq 0$ and $s \neq 0$. Using (4.2) with (4.21) gives

$$a_{0,0} = A_0 - \{A_{-1} - a_{-1,-1} - \sum_{r=1}^{4} a_{-1,r}\} - \sum_{n=1}^{4} \{A_n - a_{n,-1} - \sum_{r=1}^{4} a_{n,r}\} . \tag{4.22}$$

So also $a_{0,0}$ is expressed in terms of $a_{r,s}$ with both $r \neq 0$ and $s \neq 0$.

If we add and subtract (4.14) and (4.16) we will get

$$a_{-1,m} = \frac{1}{2} \{C_m - B_m - \sum_{n=2}^{4} n(n-1)a_{m,n}\} \tag{4.23}$$

$$a_{1,m} = \frac{1}{2} \{C_m + B_m - \sum_{n=2}^{4} n(n+1)a_{m,n}\} . \tag{4.24}$$

Except for $m = -1$, 0, and 1, (4.23) expresses $a_{-1,m}$ (and hence $a_{m,-1}$) in terms of $a_{r,s}$ with $r \geq 2$ and $s \geq 2$. If we take $m = -1$ in (4.23), and make another use of (4.23), we get

$$a_{-1,-1} = \frac{1}{2} \{C_{-1} - B_{-1} - \frac{1}{2} \sum_{n=2}^{4} n(n-1)[C_n - B_n - \sum_{r=2}^{4} r(r-1)a_{n,r}]\} . \tag{4.25}$$

If we take $m = 1$ in (4.23), and make a use of (4.24), we get

$$a_{-1,1} = \frac{1}{2} \{C_1 - B_1 - \frac{1}{2} \sum_{n=2}^{4} n(n-1)[C_n + B_n - \sum_{r=2}^{4} r(r+1)a_{n,r}]\} . \tag{4.26}$$

So, except for $m = 0$, we have $a_{-1,m}$ (and hence $a_{m,-1}$) in terms of $a_{r,s}$ with $r \geq 2$ and $s \geq 2$. Using these in (4.21) gives also $a_{-1,0}$ in terms of $a_{r,s}$ with $r \geq 2$ and $s \geq 2$.

Except for $m = 0$ and 1, (4.24) or (4.26) expresses $a_{1,m}$ in terms of $a_{r,s}$ with $r \geq 2$, and $s \geq 2$. If we take $m = 1$ in (4.24), and make another use of (4.24), we get

$$(4.27)\quad a_{1,1} = \frac{1}{2}\{C_1 + B_1 - \frac{1}{2}\sum_{n=2}^{4} n(n+1)[C_n + B_n - \sum_{r=2}^{4} r(r+1)a_{n,r}]\}.$$

If we use this, (4.26), and (4.24) in (4.21) we get also $a_{1,0}$ in terms of $a_{r,s}$ with $r \geq 2$ and $s \geq 2$.

In view of (4.2), there are only six distinct parameters $a_{r,s}$ with $r \geq 2$ and $s \geq 2$. There remain yet unused three of the eighteen original equations. It would be expected that they would give three more conditions among the $a_{r,s}$, but surprisingly they turn out to be dependent on the other fifteen. This is due to the particular relations that subsist among the A_m, B_m, and C_m, and would not be the case with general A_m, B_m, and C_m.

Thus consider (4.23) for $m = 0$, of which we have not yet made any use. If we substitute from (4.23) into (4.21), we will get

$$(4.28)\quad a_{-1,0} = A_{-1} - \frac{1}{2}\{C_{-1} - B_{-1} - \sum_{n=2}^{4} n(n-1)a_{-1,n} + \sum_{n=1}^{4}(C_n - B_n) - \sum_{n=1}^{4}\sum_{r=2}^{4} r(r-1)a_{n,r}\}.$$

Making use of $a_{n,r} = a_{r,n}$ lets us write the final term as

$$\sum_{n=2}^{4} n(n-1) \sum_{r=1}^{4} a_{n,r} .$$

By (4.15) and (4.17), we have

$$C_{-1} - B_1 + \sum_{n=1}^{4} (C_n - B_n) = 1 - (C_0 - B_0) .$$

Also, from the given values of the A_m, we have

$$A_{-1} = \tfrac{1}{2} \{1 - \sum_{n=2}^{4} n(n-1)A_n\} .$$

Putting these into (4.28) gives

$$a_{-1,0} = \tfrac{1}{2} \{C_0 - B_0 - \sum_{n=2}^{4} n(n-1)[A_n - a_{-1,n} - \sum_{r=1}^{4} a_{n,r}]\} .$$

Use of (4.21) converts this into (4.23) with $m = 0$.

Consider next (4.24) for $m = 0$, of which we have not yet made any use. If we substitute from (4.24) into (4.21), we will get

$$(4.29) \quad a_{1,0} = A_1 - \tfrac{1}{2} \{C_{-1} + B_{-1} - \sum_{n=2}^{4} n(n+1)a_{-1,n} + \sum_{n=1}^{4} (C_n + B_n) - \sum_{n=1}^{4} \sum_{r=2}^{4} r(r+1)a_{n,r}\} .$$

As before, we write the last term as

$$\sum_{n=2}^{4} n(n+1) \sum_{r=1}^{4} a_{n,r} ,$$

and we have

$$C_{-1} + B_{-1} + \sum_{n=1}^{4} (C_n + B_n) = 1 - (C_0 + B_0) ,$$

and

$$A_1 = \frac{1}{2} \{1 - \sum_{n=2}^{4} n(n+1)A_n\} .$$

Substituting these into (4.29), and using (4.21), gives (4.24) with $m = 0$.

Consider finally (4.24) with $m = -1$. Refer back to (4.26). We have

$$\frac{1}{4} \sum_{n=2}^{4} n(n-1) \sum_{r=2}^{4} r(r+1)a_{n,r}$$

$$= \frac{1}{4} \sum_{n=2}^{4} n(n+1) \sum_{r=2}^{4} r(r-1)a_{n,r} .$$

Also, use of (4.15) and (4.18) gives

$$\frac{1}{2} \{C_1 - B_1 - \frac{1}{2} \sum_{n=2}^{4} n(n-1)(C_n + B_n)\}$$

$$= \frac{1}{2} \{C_{-1} + B_{-1} - \frac{1}{2} \sum_{n=2}^{4} n(n+1)(C_n - B_n)\} .$$

Substituting these into (4.26) and using (4.23) gives (4.24) with $m = -1$.

Thus we can choose $a_{r,s}$ with $r \geq 2$ and $s \geq 2$ at will, subject to $a_{r,s} = a_{s,r}$. Then we can substitute gradually back, and recover all

the $a_{r,s}$. Recalling that we have also the free parameters B_0, C_0, and C_4, we see that there is a nine fold multiplicity of solutions. One would have thought it possible to choose $a_{4,n} = 0$ for all n, thus reducing to the case $S = 3$. However, as $A_4 = 1/40$, this is precluded by (4.6).

We observe that if the principal grid point is at a distance h from the left edge, but further than that from the top or bottom, then one can use the off center difference approximation in the x-direction only. We use the methods given above to see if one can get a stencil which does not extend as far as six grid points in the x-direction. It turns out that one cannot, but we will present the analysis anyhow, since it shows how to generate all possible stencils.

Without causing confusion, we can use the same letters as before, but with slightly altered denotations.

So for our $a_{m,n}$ we will now have $-2 \leq n \leq 2$, $-1 \leq m \leq S$. In place of (4.2), we will have

$$a_{m,n} = a_{m,-n}. \tag{4.30}$$

All summations on n should be from -2 to $+2$. Specifically, this change should be made in (4.3), (4.4), (4.6), (4.14), and (4.16). Delete (4.5).

As before, we see that we must have $S \geq 4$. Taking $S = 4$, we get the same values of A_m as before.

By (4.30) and (4.14), we have $B_m = 0$ for all m. Thus (4.15) is trivially satisfied.

We get the same determination as before for the C_m.

Finally, we write

$$D_m = \sum_{n=-2}^{2} n^4 a_{m,n} . \tag{4.31}$$

We must have

$$D_{-1} + D_0 + D_1 + D_2 + D_3 + D_4 = \frac{2}{5} \tag{4.32}$$

$$-D_{-1} + D_1 + 2D_2 + 3D_3 + 4D_4 = 0 . \tag{4.33}$$

Given the A_m, B_m, C_m, and D_m, there is no question how to determine the $a_{m,n}$. We have immediately

$$a_{m,-2} = a_{m,2} = \frac{D_m - C_m}{24} ,$$

$$a_{m,-1} = a_{m,1} = \frac{4C_m - D_m}{6} ,$$

$$a_{m,0} = A_m - 2a_{m,1} - 2a_{m,2} .$$

Thus we can easily determine sets of $a_{m,n}$. There are 18 distinct $a_{m,n}$. As they do not depend on the B_m, it appears that we have a six fold multiplicity of solutions. It is surprising that this does not permit the choice $a_{4,0} = a_{4,1} = a_{4,2}$, which would let us reduce S to 3.

5. Regions of unusual shape. We have been using squares for our grid. There are cases where this is really impractical. For example, suppose our region is a rectangle of sides 1 and $\sqrt{2}$. For rectangles of intractable proportions, a way of handling the matter easily is provided in Rosser [1]. Beyond that, we have not pushed our investigations.

REFERENCES

[1] J. Barkley Rosser, Finite-difference solution of Poisson's equation in rectangles of arbitrary shape, Technical Report TR/27, Brunel University, 1974, and MRC Technical Summary Report #1404, February 1974.

[2] George E. Forsythe and Wolfgang R. Wasow, Finite-difference methods for partial differential equations, John Wiley and Sons, Inc., New York, 1960.

[3] L. Collatz, The numerical treatment of differential equations, Springer-Verlag, Berlin, 1966.

ILLUMINATING ROUND EFFECTIVENESS MODELING

Martin Messinger
and
Leonard Oleniczak

Picatinny Arsenal
Ammunition Development and Engineering Directorate
Dover, New Jersey

Abstract

This paper is concerned with the problem of evaluating the effectiveness of existing and conceptual illuminating rounds. The illuminating round effectiveness model developed at Picatinny Arsenal essentially consists of three parts: a ballistic portion which is concerned with round deployment and descent, an illumination/atmospheric portion which calculates target/background visual contrasts at all points in search area during the entire flare burn time, and a human visual perception portion which determines from the target/background contrast the glance target detection and/or recognition probabilities. Effectiveness measures are evaluated by integrating suitable functions of the glance target detection/recognition probabilities over the search area and flare burn time.

In particular, this paper will concentrate on developing the illuminating round effectiveness measures built into the model. Basically, two different measures of effectiveness are employed. The first measure, referred to as the time integral of the area illuminated (TIAI), has the dimensions of area-time and is meaningful when it is desired to evaluate the ability of an illuminating round to maintain illumination over a given search area. The second measure of effectiveness, referred to as effective area (EA), has the dimension of area and is, in spirit, analogous to the lethal area concept employed extensively in evaluating the effectiveness of HE munitions. This effectiveness measure is particularly useful when it is desired to evaluate the ability of an illuminating round to enable detection or recognition of unknown targets in a sepcified search area.

An application of the model to an hypothetical incremental flare illuminating round is included to illustrate concepts developed.

Survey of Illumination/Atmospheric and Visual Perception Sub-Models

The scenario employed in the model is illustrated in Fig. 1. Though the figure depicts, as an example, a ground launched round, the

model can just as readily consider air dropped flares. The model assumes a flat rectangular search area which is to be illuminated by the round. Inputs to the model include parameters to describe the characteristics of the flare, parameters to define the scenario, parameters to specify target/background characteristics, and parameters to characterize the atmosphere. Table 1 gives a detailed list of the inputs necessary to drive the model.

TABLE 1

INPUTS TO THE MODEL

Flare Parameters:

- Candle Intensity Vs Time
- Burn Rate Vs Time
- Pyrotechnic Weight
- Non-Combustible Weight
- Initial Stabilized Flare Descent Rate

Scenario Parameters:

- Search Area Dimensions and Location
- Location of Observers
- Number of Observers
- Observer Scan Rate
- Initial Flare Position

Target/Background Parameters:

- Target Size
- Target Reflectivity
- Background Reflectivity

Atmospheric Parameters:

- Normalized Volume Scattering Function
- Meterological Range (Visibility)
- Wind

As shown in Figure 2, targets in search area are modeled as having two reflecting surfaces; a horizontal surface parallel to the ground plane and a vertical surface oriented such that its normal is pointed in the direction of the projection of the observer's location into the ground plane. Target/background contrasts and detection/recognition probabilities are calculated independently for

each surface.

The illumination geometry is portrayed in Figure 3. Light reaching the observer comes from three sources: light reflected from the target surface as the result of direct illumination by the flare, light from the flare scattered into the visual path between the target and the observer which gives rise to path luminance, and glare which essentially represents direct light from the flare reaching the observers eye. The apparent target/background contrast is defined as:

$$C = \frac{\left| B_T - B_B \right|}{B_B} \qquad (1)$$

where C = apparent target/background contrast as seen by the observers.

B_T = apparent target brightness as seen by the observers (footlamberts).

B_B = apparent background brightness as seen by the observers (footlamberts).

In calculating the apparent brightness of the target and background it is, of course, necessary to account for the light attenuation that results from atmospheric scattering (light absorption in the visible spectrum is insignificant and is neglected). It is also important to note that the model accounts for only single scattering and neglects indirect illumination and other effects due to multiply scattered light.

From the apparent target/background contrast one computes the target detection probability. This computation is performed with the aid of the extensive work performed by Blackwell.[1,2] Using Blackwell's results, the detectability at time t of a target located at (x, y) in the search area (i.e., the conditional target detection probability given the presence of a target and a observer looking at the location) is given by:

$$P(x,y,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{[C(x,y,t)/C_L(x,y)-1]/.39} e^{-u^2/2}\, du \qquad (2)$$

where $C(x, y, t)$ = the actual apparent target/background contrast at time t for a target located at (x, y).

$C_L(x, y)$ = the liminal target/background contrast, (i.e., contrast required for 50% target detection probability) for a target located at (x, y).

In general, the liminal target/background contrast is a function of the visual angle subtended by the target to the source, the background brightness which determines the visual adaptation level, and the time duration that the location is examined. Fig. 4 gives the liminal target/background contrast used in the study. It is based on a 1/3 second observer glance duration.

The probability of target recognition given detection, P_r, depends upon the additional ability of the observer to resolve individual portions of the target so that its shape can be ascertained. For target detection, all that is required is for the observer to be able to identify an anomaly in the terrain features. The expression used for the conditional probability of recognition given target detection is taken from the MARSAM II[3] study:

$$P_r(x,y,t) = \begin{cases} 1-\exp\dfrac{(N(x,y,t)-3.2)^2}{11} & N \geqq 3.2 \\ 0 & N < 3.2 \end{cases} \qquad (3)$$

where N (x,y,t) denotes the number of resolvable elements across the target surface. N(x,y,t) depends upon the angle subtended by the target to the observer and visual adaptation level. The product $P(x,y,t)\cdot P_r(x,y,t)$ gives the unconditional glance probability of target recognition.

Target detection and/or recognition glance probabilities are computed independently for both the horizontal and vertical target surfaces. Depending upon the intended application, either the horizontal surface detection or recognition, vertical surface detection or recognition or maximum detection or maximum recognition probabilities can be used in computing the illuminating round effectiveness measures. For simplicity, in the derivations that follow we will use one of the glance target detection probabilities, P(x,y,t).

Measures of Effectiveness

The first effectiveness measure to be considered is the time integral of the area illuminated, TIAI. At a given time, t, during the flare burn, one can weight each differential area, dA, in the search region by the target detectability, P(x,y,t), and integrate over the

entire search area to obtain the area illuminated at time t, AL(t).

$$AL(t) = \int_{\text{Search Area}} P(x,y,t)dA \qquad (4)$$

The significance of Eq. 4 becomes apparent if one envisions the search area to be populated with targets of uniform area density ρ targets/m^2. Thus $\rho\, P(x,y,t)dA$ is the expected number of detectable targets in dA and hence, the expected number of detectable targets in the search area at time t is given by $\rho\, AL(t)$.

Integrating Eq. 4 over the burn time of the flare gives rise to the time integral of the area illuminated.

$$TIAI = \int_{\text{Burn Time}} \left\{ \int_{\text{Search Area}} P(x,y,t)dA \right\} dt \qquad (5)$$

The maximum possible value for TIAI is the product of the search area and the flare burn time. Normalizing by this quantity yields an effectiveness measure whose value lies between 0 and 1 and represents the average,over search area and flare burn time, of the target detectability. This figure of merit is most useful in evaluating the ability of a round to maintain adequate illumination during the burn over the entire search area.

$$\text{Average detectability} = \frac{TIAI}{AT} \qquad (6)$$

In order to develop the notion of the Effective Area, EA, of an illuminating round in a given scenario, the concept of the observer scan rate must be introduced. The scan rate denoted by α , is simply the area scanned by each observer per unit time. We assume that the scanning procedure is random and that the observers are independent. Hence in differential time, Δt, the probability that a given point (x,y) in the target area is examined is given by:

$$\text{Prob. (glimpse at } (x,y) \text{ in } \Delta t) = \frac{n\alpha\Delta t}{A} \qquad (7)$$

The probability of detecting a target at (x,y) not previously detected in the time interval t to $t + \Delta t$ is thus given by:

$$\text{Prob}\left[\begin{array}{l}\text{Target at } (x,y) \\ \text{detected between} \\ t, t+\Delta t\end{array}\Big/\quad \begin{array}{l}\text{Not detected} \\ \text{up to time t}\end{array}\right] = P(x,y,t)\,\frac{n\alpha\Delta t}{A} \qquad (8)$$

Let $K(x,y,t)$ denote the probability of detecting a target located at coordinates (x,y) in the time interval 0 to t, and let $\bar{K}(x,y,t) = 1 - K(x,y,t)$ denote the probability of not detecting the target by time t. We then obtain the following difference equation for $\bar{K}(x,y,t)$:

$$\bar{K}(x,y\ t+\Delta t) = \bar{K}(x,y,t) \quad [\ 1-P\ (x,y,t)\ \frac{n\alpha\Delta t}{A}\ ] \tag{9}$$

This difference equation results from the fact that for a target not to be detected by $t + \Delta t$, it must not be detected up to t and not detected in the interval t to $t + \Delta t$. Taking the limit as $\Delta t \to 0$, one obtains the differential equation:

$$\frac{d}{dt}\ \bar{K}(x,y,t) = -P\ (x,y,t)\ \frac{n\alpha}{A}\ \bar{K}(x,y,t) \tag{10}$$

The solution to this equation subject to the initial condition that at $t = 0$, $\bar{K}(x,y,t) = 1$ is given by:

$$\bar{K}(x,y,t) = e^{-\frac{n\alpha}{A}\int_0^t P(x,y,\eta)\,d\eta} \tag{11}$$

Hence the probability of detecting a target located at coordinates (x,y) over the total flare burn time T is given by

$$K(x,y,T) = 1 - e^{-\frac{n\alpha}{A}\int_0^T P(x,y,t)\,dt} \tag{12}$$

Finally, the probability of detecting a target in dA over the flare burn is given by the product of the probability that there is a target in dA and the conditional detection probability given the target is in dA.

$$K(x,y,T)\ \rho dA \tag{13}$$

The expected number of targets detected is thus given by

$$\text{Expected number of targets detected} = \rho\int_A [1-e^{-\frac{n\alpha}{A}\int_0^T P(x,y,t)\,dt}]\ dA \tag{14}$$

The actual number of target detection would be Poisson with the above expectation.

As an effectiveness measure one can thus take the expression

$$EA = \int_A [1-e^{-n\alpha/A \int_0^T P(x,y,t)dt}] \, dA \qquad (15)$$

The above quantity referred to as the Effective Area, has the dimension of area. Its value is between 0 and the size of the area being illuminated. One can define a flare efficiency by considering the ratio of EA to either the total search area or the expected area scanned by the observers during the search. The first normalization yields:

$$\frac{EA}{A} = \frac{\int_A [1-e^{-n\alpha/A \int_0^T P(x,y,t)dt}]dA}{A} \qquad (16)$$

For a uniform target area density, this quantity can be interpreted as the ratio of the expected number of targets present in the search area.

An alternate interpretation is possible for the case where it is desired to find a single target located in the search area. The conditional probability that the target is detected given that it is located at (x,y) is given from Eq. 12 as:

$$\left\{ \begin{matrix} \text{Prob. of detecting} \\ \text{target at } (x,y) \end{matrix} \right\} = 1 - e^{-\frac{n\alpha}{A}\int_0^T P(x,y,\eta)d\eta} \qquad (17)$$

The probability that the target is in dA is, assuming all locations are equally likely, clearly given by dA/A. Hence integrating over the search area demonstrates that EA/A also represents the probability of finding a single target located randomly in the search area.

A potential difficulty with Eq. 16 is that it represents the combined capabilities of the observers as well as the flare. From Eq. 15, one can obtain the expected total area scanned by all the observers by setting P(x,y,t) equal to 1

$$\text{Expected Area Scanned} = A[1-e^{-\frac{n\alpha T}{A}}] \qquad (18)$$

If the expected total scanned area is small, EA/A will be small no matter how good the actual flare performance. To account for the

possible limitation of the observers to scan the entire search area, one can evaluate, for the single target case, the conditional probability that the target is detected given that it is scanned. This probability is given by:

$$P\ (\text{target detected/target scanned}) = \frac{P\ (\text{target detected})}{P\ (\text{target scanned})} \tag{19}$$

The probability that the target is detected is given by EA/A, and the probability that the target is scanned is obtained from Eq. 11 by setting P(x,y,t) equal to 1. Performing these substitutions we obtain:

$$P\ (\text{target detection/target scanned}) = \frac{EA}{A\left(1 - e^{-\frac{n\alpha T}{A}}\right)} = \frac{\text{Effective Area}}{\text{Expected Area Scanned}} \tag{20}$$

The effectiveness expressions obtained in this section are illustrated below with the aid of an example.

Example

Consider a Hypothetical Incremental Flare whose candle power and descent altitude-time profile is shown in Fig. 5. The scenario consists of a search area which is a square 3000 feet on each side and a single observer who can scan 10^6 ft^2/sec., located at an altitude of 200 feet, 6500 feet back from the center of the search area. The reflectivity of the target is 0.14 and the background is 0.10. The target diameter is 15 feet. The flare is dropped directly over the center of the search area, wind drift is assumed to be zero, and the atmospheric visibility is 15 miles.

Figure 6 shows the glance target detection probability that exists 30 seconds after flare initiation. Notice the rapid fall off in glance detection probability on the front lighted portions of the search area. Figure 7 depicts the area illuminated, Eq. 4, as a function of flare burn time, and Figure 8 depicts the effectiveness measures as a function of flare burn time. Curve A indicates that the maximum average target detectability occurs early in the flare burn. It may be desirable to design the candle to provide a high target detectability at the beginning of its burn before countermeasure actions can be taken by the enemy. A flat curve would indicate a flare design which provides constant average target detectability during its entire burn. Curves B and C show the probabilities of single target detection up to a given time. The two curves converge since, for the parameters used, the probability of

scanning the entire search area during the entire flare burn is virtually 1.

Conclusions

The model described in this paper is a general purpose system model developed at Picatinny Arsenal for the purpose of evaluating illuminating round effectiveness. The model has already been successfully applied to answer questions regarding fuze accuracy requirements. The model should be particularly applicable in evaluating advanced illuminating round design concepts such as the incremental flare. In fact the model can be used to evaluate the relative importance of each of the design parameters and thus enable intelligent decisions regarding the allocation of development funds. The model can also be applied to questions of doctrine, tactics and the development of firing tables.

Comprehensive documentation of the model, including the computer code, is given in Reference 4.

References

1. Blackwell, Richard H., "Contrast Thresholds for the Human Eye", Journal of the Optical Society of America, Volume 36, Number 11, November 1946.

2. Blackwell, Richard H., "Development and Use of a Quantitative Method for Specification of Interior Illumination levels on the Basis of Performance data,"Illuminating Engineering,June 1959.

3. Greening, Charles P., "Target Acquisition Model Evaluation Final Summary Report", Technical Report Number NWC TP5536, Naval Weapons Center, China Lake, Calif., June 1973.

4. Messinger, M., Oleniczak L., "Illuminating Round Effectiveness Model", Picatinny Arsenal Technical Report.

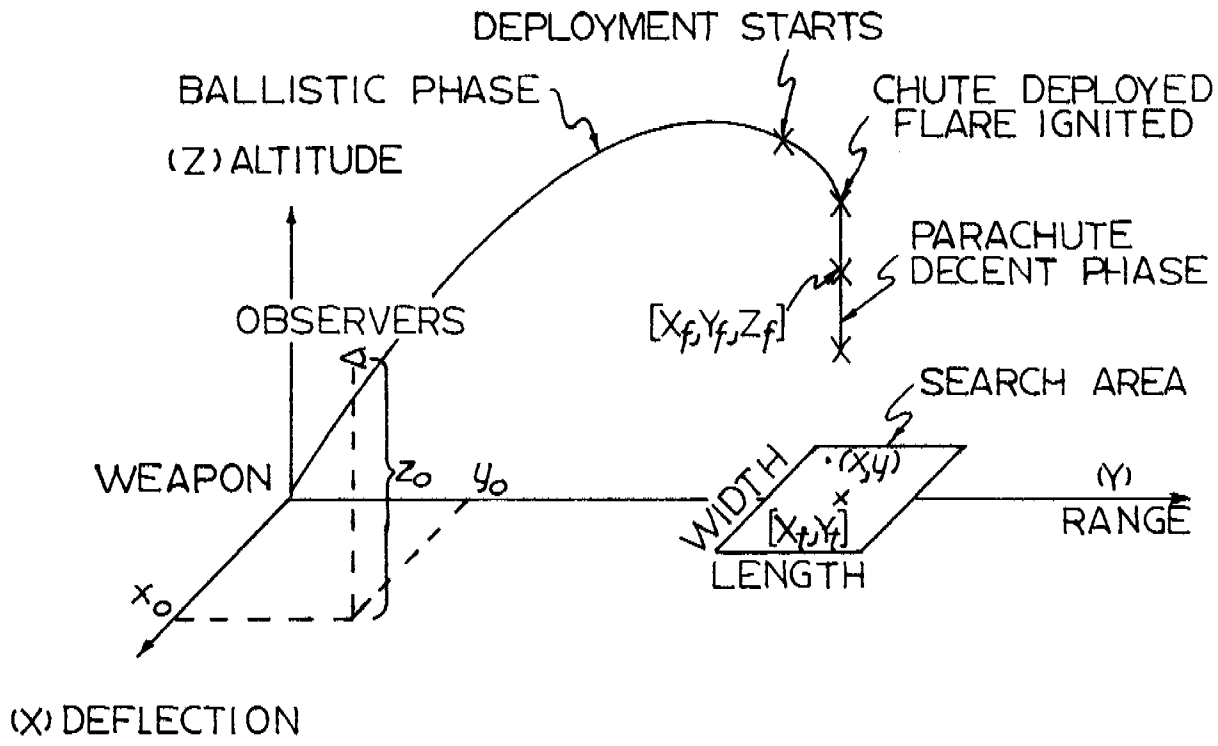
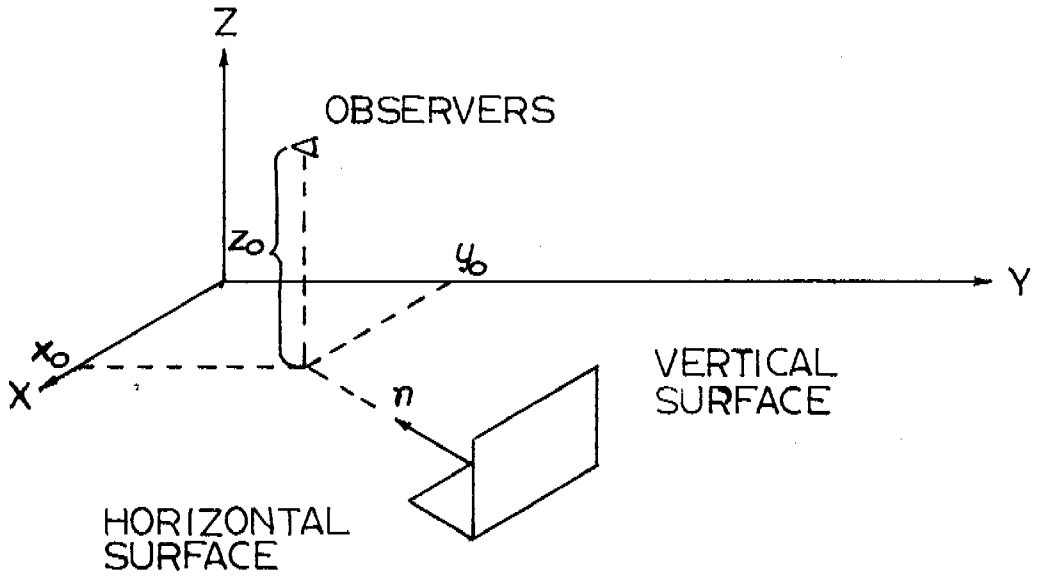
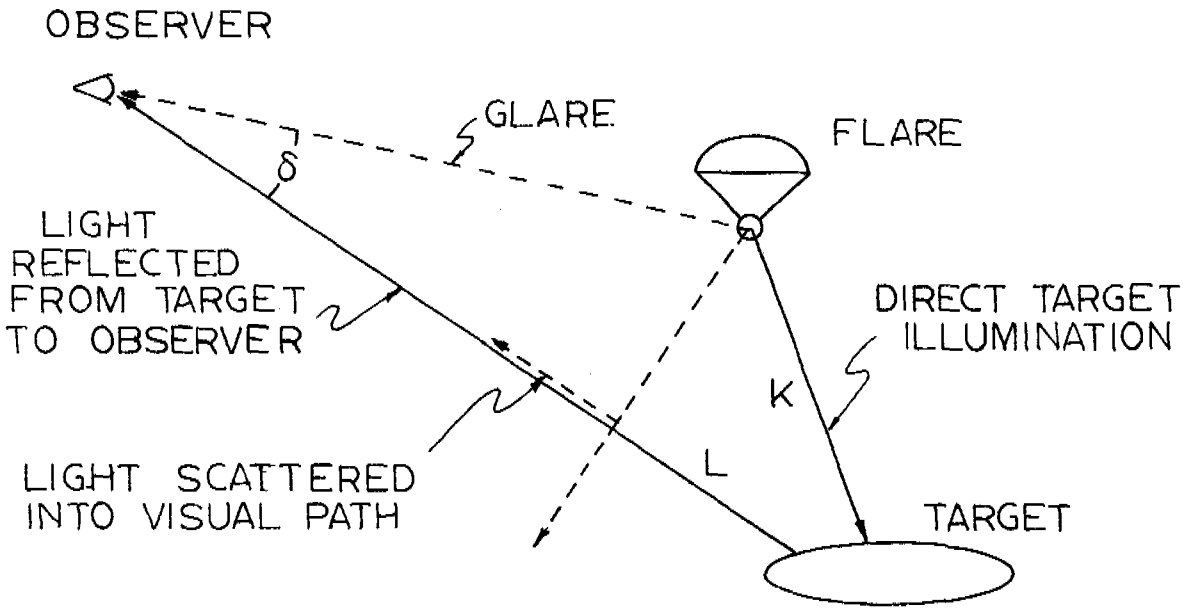


fig. 1 SCENARIO



TARGET GEOMETRY

fig. 2



ILLUMINATION GEOMETRY

fig. 3

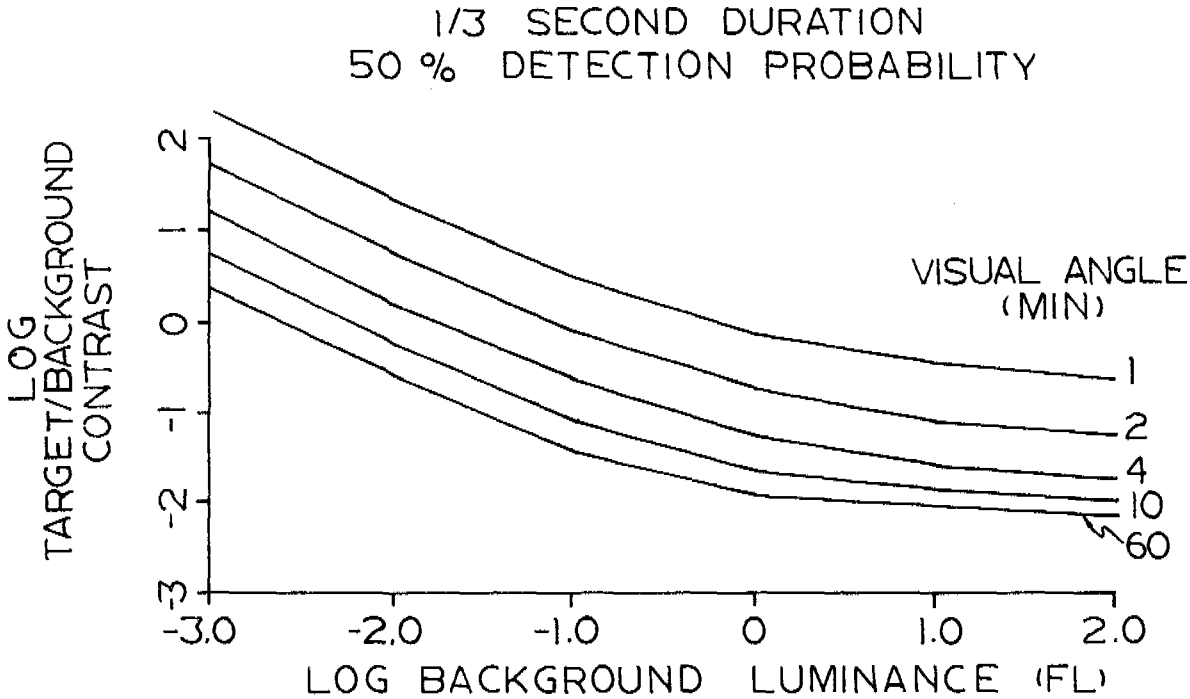


fig. 4

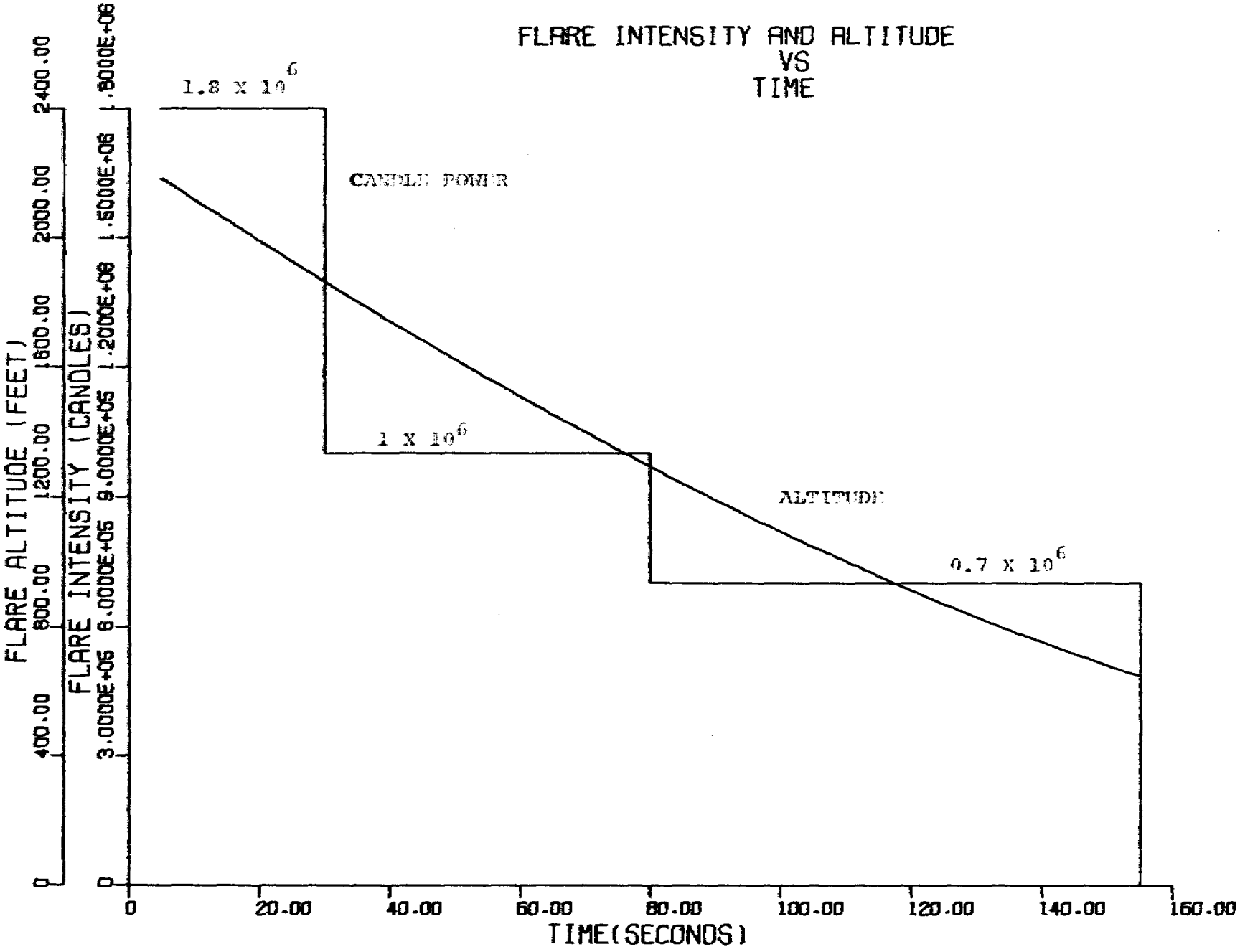


FIGURE 5: CANDLE INTENSITY AND ALTITUDE PROFILE FOR INCREMENTAL FLARE

PA LONG INCREMENT A
16.28.20. 04/11/75
SEC. AFTER FLARE INIT.= 30.00
RT= .14 RB= .10 FLARE CDL PWR= 1800000.
OBS. LOC. (FT) XO= 0. YO= 0. ZO= 200.
FLARE LOC. (FT) XF= 0. YF= 6500. ZF= 1871.
TARGET DIAMETER (FT)= 15.00
TARGET LOC (FT) XC= 0. YC= 6500. DX= 3000. DY= 3000.
VISIBILITY (METERS)=24160.0
CLOUD BASE HEIGHT=INFINITE
VERTICAL TARGET ONLY

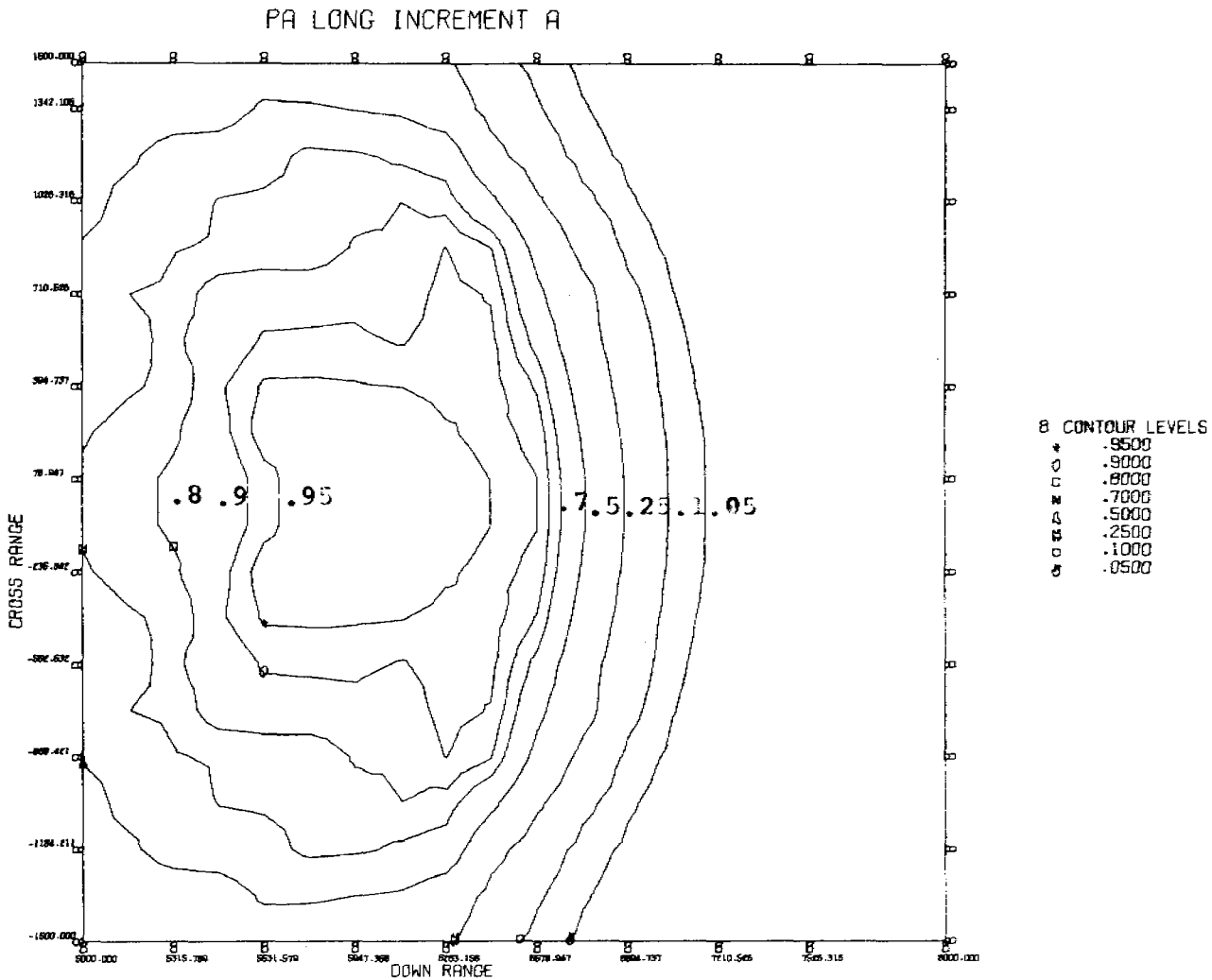


FIGURE 6: DETECTABILITY CONTOURS AT 30 SECONDS INTO FLARE BURN

LAUNCHING OF ELECTROMAGNETIC SURFACE WAVES AT A PLANAR METAL-AIR INTERFACE

J. M. Zavada and E. L. Church
Pitman-Dunn Laboratory
Frankford Arsenal
Philadelphia, PA 19137

ABSTRACT. We consider the two-dimensional problem of a line source located in air above a homogeneous metal half space. The incident field is represented as an angular spectrum of plane waves, and integral expressions for the reflected and transmitted fields are obtained by satisfying the usual electromagnetic boundary conditions at the interface for each plane wave. The properties of these fields are examined in the far-field region by evaluating the integrals by the method of steepest descents, including the effects of branch cuts and poles. This method indicates that for certain values of the complex permittivity of the metal, slow Zenneck surface waves are excited. These waves decay exponentially away from the interface, propagate along the surface with a phase velocity less than that of light, and have a transverse attenuation length determined by the conductivity of the surface. An expression is derived for the power carried off by these surface waves.

1. INTRODUCTION. In recent years there has been considerable interest in the generation and propagation of electromagnetic surface waves at optical and infrared frequencies. These waves are characterized by an exponential decay away from the guiding surface and have been excited on a number of open-boundary structures.[1] In particular, these surface waves have been generated directly on corrugated or dielectric coated planes; and indirectly, by means of a prism coupler, on smooth, uncoated plane surfaces.[2] The main interest in these waves lies in their potential application as surface probes[3] and in various electronic devices.[4]

These surface waves may also play a significant role in the increased absorption of mirrors due to surface microroughness.[5] To determine their actual importance, it is necessary to calculate the amount of energy that is coupled into these waves by surface corrugations or roughness. Our approach to this problem has been based on a classical EM scattering formalism which was developed for radar scattering, augmented with a perturbation expansion appropriate to small scales of roughness.[6]

In our investigation of the effects of surface roughness, we have found that electromagnetic surface waves can be excited on a smooth, uncoated plane surface without the use of a prism coupler.[7] The present paper demonstrates this result and specifies the conditions which are necessary for the launching of these waves. This paper also outlines the general formalism used in our continuing studies of slightly rough surfaces.

2. PLANE WAVE REPRESENTATION OF FIELDS. We treat the two-dimensional problem of a line source situated in air a distance h above a homogeneous half plane (Fig. 1). The permeability of the entire plane is that of the vacuum, μ_o. The permittivity of air is ϵ_o (the free space value) and that of the medium is

$\bar{\epsilon} = (\epsilon_1 - i\epsilon_2)\,\epsilon_o = \epsilon\epsilon_o$. We allow $\bar{\epsilon}$ to be a complex quantity to allow for possible losses in the medium. The line source has a time dependence $e^{i\omega t}$ (which will be suppressed in the following) and is specified by:

$$H_z^i(x,y) = \sqrt{\frac{\pi}{2}} \exp\left(-\frac{i\pi}{4}\right) H_o^{(2)}(k_o r') \tag{1}$$

$$H_x^i = H_y^i = 0$$

where (x,y) are the coordinates of the observation point; r' is the distance from the source; $k_o = 2\pi/\lambda$ and $H_o^{(2)}(..)$ is the zeroth order Hankel function of the second kind. We have chosen a p-polarized line source since only this polarization will lead to surface-wave excitation in this particular problem.

The first step in our calculation is to represent the incident fields in terms of a spectrum of plane waves.[8] This procedure is closely related to the Fourier expansion of the fields and originates in the work of Sommerfeld, Debye and Weyl. In fact, one of the standard integral representations of the Hankel function is in the form of a plane wave expansion:

$$H_o^{(2)}(x,y) = \frac{1}{\pi}\int_C \exp(+ i k_o(x\cos\alpha \mp y \sin\alpha))\, d\alpha \tag{2}$$

where the sign convention (- for $y > 0$, + for $y < 0$) is chosen to guarantee convergence of the integral. In Eqn. (2), α is a complex angle and C is the path of integration in the α plane (Fig. 2). Then, selecting the representation which is valid for the half-space below the source, we obtain the plane wave expansion for H_z^i:

$$H_z^i(x,y) = \frac{K}{\pi}\int_C P(\alpha) \exp(i k_o(x\cos\alpha + y \sin\alpha))\, d\alpha \tag{3}$$

where $K = \sqrt{\frac{\pi}{2}} \exp\left(-\frac{i\pi}{4}\right)$ and $P(\alpha) = \exp(- i k_o h \sin\alpha)$. Here $P(\alpha)$ is the spectral density of the field and the exponential factor is a plane wave. Since α can be a complex quantity, we have two types of plane waves in the spectrum. When α is real, we have the usual homogeneous plane wave of constant amplitude. When α is a complex angle, we have an inhomogeneous wave, i.e., the planes of constant phase no longer coincide with the planes of constant amplitude.[9]

In order to obtain the reflected and transmitted fields, we satisfy the usual EM boundary conditions at the surface for each plane wave of the spectrum. This leads to the following representation of the reflected field H_z^R:

(4) $$H_z^R(x,y) = \frac{K}{\pi}\int_c \rho_H(\alpha)\, P(\alpha)\, \exp(i\, k_o(x\cos\alpha - y\sin\alpha))\, d\alpha$$

$$\rho_H(\alpha) = \frac{\epsilon\sin\alpha - \sqrt{\epsilon - \cos^2\alpha}}{\epsilon\sin\alpha + \sqrt{\epsilon - \cos^2\alpha}}$$

where $\rho_H(\alpha)$ is the Fresnel reflection coefficient for a p-polarized plane wave. Similarly, the transmitted field H_z^T is represented by:

(5) $$H_z^T(x,y) = \frac{K}{\pi}\int_c \tau_H(\alpha)\, P(\alpha)\, \exp(i\, k_o(x\cos\alpha + y\sqrt{\epsilon - \cos^2\alpha}))\, d\alpha$$

$$\tau_H(\alpha) = 1 + \rho_H(\alpha)$$

where $\tau_H(\alpha)$ is the Fresnel transmission coefficient for p-polarization.

3. SADDLE POINT INTEGRATION. Since the integral expressions for H_z^R and H_z^T cannot be solved in closed form, we investigate the properties of these fields by the method of steepest descents. In this method, the original path of integration in the α plane is deformed into the path of steepest descents,[10] and then an asymptotic expansion is performed to obtain the far-field behavior of the fields. However, if there are any singularities in the integrand (poles or branch points), Cauchy's theorem requires that we include their contribution to the integral. Deforming the contour C into the path of steepest descents S, then leads to the general result:

(6) $$\int_c = \int_S + \int_{b.c.} + 2\pi i \sum \text{Res.}$$

In Eqn. (6) $\int_{b.c.}$ represents the integration along the branch cuts which are crossed and $\sum$ Res. is the contribution from the residues of the integrand at the poles which are encountered. The integral along S, $\int_S$, yields the radiation field and $\int_{b.c.}$ leads to lateral waves.[11] The residues represent the surface waves which are excited. Therefore, the saddle point integration gives us a means of distinguishing the surface waves in the resulting fields. Also we notice that the fields will contain surface waves only if the integrand has a pole which is encountered as the contour is deformed.

First, we apply this method to the reflected field H_z^R. It is convenient at this stage to introduce polar coordinates (R, ψ) defined as in Fig. 1:

(7) $$R\cos\psi = x$$
$$R\sin\psi = y + h$$

Then Eqn. (3) can be rewritten:

(8) $$H_z^R = \frac{K}{\pi}\int_c \rho_H(\alpha)\, \exp(i\, k_o\, R\cos(\psi + \alpha))\, d\alpha$$

The saddle point is easily found to be $\alpha_S = \pi - \psi$ and the curve of steepest descent passing through α_S is defined:

$$\begin{aligned} \cos\ (\sigma - \alpha_S)\ \cosh\ \upsilon &= 1 \\ \sin\ (\sigma - \alpha_S &= \tanh\ \upsilon \end{aligned} \tag{9}$$

where $\sigma = \text{Re}\ \alpha$ and $\upsilon = \text{Im}\ \alpha$. This curve is denoted by $S(\alpha_S)$ and is shown in Fig. 3.

The only poles in the integrand of Eqn. (8) are those of $\rho_H(\alpha)$.[12] From Eqn. (4), we find that these poles are defined by:

$$\begin{aligned} \cos\ \alpha_\rho &= \pm\sqrt{\frac{\epsilon}{1+\epsilon}} \\ \sin\ \alpha_\rho &= -\frac{1}{\sqrt{1+\epsilon}} \end{aligned} \tag{10}$$

Two of these poles are relevant to the present discussion. If we let $\sqrt{\epsilon/1+\epsilon} = A - iB$ and $1/\sqrt{1+\epsilon} = D_1 + iD_2$, then these two poles α_o and α_π are given by:

$$\begin{aligned} \alpha_o &= -\sigma_\rho - i\ \upsilon_\rho \\ \alpha_\pi &= \pi + \sigma_\rho + i\ \upsilon_\rho \\ \tan\ \sigma_\rho &= \frac{D_1}{A} \\ \tanh\ \upsilon_\rho &= \frac{D_2}{A} \end{aligned} \tag{11}$$

The approximate location of these poles in the α plane is shown in Fig. 3.

The coefficient $\rho_H(\alpha)$ also has branch point singularities α_B which are defined:

$$\cos\ \alpha_B = \pm\sqrt{\epsilon} \tag{12}$$

These branch points and the appropriate branch cuts are also displayed in Fig. 3.

Referring to Fig. 3, we have the following situation: The original contour C is deformed into the path of steepest descents $S(\pi - \psi)$ which passes through the saddle point α_S. As we vary our point of observation (ψ goes from 0 to π), we sweep out the region shaded in the figure. In doing so, we pass over branch points and this integral along the branch cuts contributes lateral waves to the field. Also, depending on the material properties of the medium (the value of ϵ), we may encounter a pole which will give rise to a surface wave.

From Eqns. (9) and (10), it can be shown that if A, the real part of $\sqrt{\epsilon/1+\epsilon}$, is greater than unity, then the poles α_o and α_π will lie in the shaded region and a surface wave will be excited. In this case there is a critical angle ψ_c which delineates the regions in which the surface wave will appear in the reflected field. For an angle ψ satisfying either $\psi < \psi_c$ or $\pi - \psi < \psi_c$, a pole will be encountered in the deformation of the contour C. Physically, this means that the surface wave is important in determining the far-field expansion of H_z^R only in regions near the surface[13] (Fig. 4).

Since the value of Re $\sqrt{\epsilon/1+\epsilon}$ is essential in determining the existence of surface waves in this problem, we restate this condition in terms of ϵ_1, and ϵ_2. By solving the equation:

$$\text{Re}\sqrt{\frac{\epsilon}{1+\epsilon}} = 1 \tag{13}$$

we obtain the curve in the complex ϵ plane shown in Fig. 5. The shaded region corresponds to values of $A > 1$. Then, if the material has dielectric response such that (ϵ_1, ϵ_2) is a point in the shaded region, surface waves will appear in the reflected field.

The transmitted field is treated in much the same way. However, since the integral contains the factor $\exp(i k_o y \sqrt{\epsilon - \cos^2\alpha})$, the saddle point integration of this field is more complicated. Nevertheless, the decomposition of H_z^T into a radiation field, a lateral wave and a surface wave remains valid. The surface wave again results from the residue of the integral at either α_o or α_π. Also, the conditions for the appearance of the surface wave in H_z^T are the same as those that apply for the reflected field.

4. SURFACE WAVES. We shall now examine the surface waves that are excited along the positive x axis ($\psi < \psi_c$) when Re $\sqrt{\epsilon/(1+\epsilon)} > 1$. By calculating the residues of the reflected and transmitted fields at $\alpha_\pi = \pi + \sigma_\rho + i\,\upsilon_\rho$, we obtain:

$$H_z'(x,y) = H_o \exp\left(-i k_o\left(\sqrt{\frac{\epsilon}{1+\epsilon}}\,x - y/\sqrt{1+\epsilon}\right)\right) \qquad x > 0,\ y > 0$$

$$H_z''(x,y) = H_o \exp\left(-i k_o\left(\sqrt{\frac{\epsilon}{1+\epsilon}}\,x - \frac{\epsilon y}{\sqrt{1+\epsilon}}\right)\right) \qquad x > 0,\ y < 0 \tag{14}$$

$$H_o = 2\sqrt{2\pi i}\,\frac{\epsilon^{3/2}}{\epsilon^2 - 1}\exp(i k_o h/\sqrt{1+\epsilon})$$

The surface wave in Eqn. (14) belongs to the general class of Zenneck surface waves which are defined by the vanishing of the reflection coefficient $\rho_H(\alpha)$.[14] The field H_z' represents an inhomogeneous plane wave incident on the surface at a complex angle $\theta' = \sigma_\rho + i\,\upsilon_\rho$. At this angle, there is no reflected wave only the transmitted inhomogeneous plane wave H_z''. It is impossible to excite these surface waves on a smooth, uncoated surface with incident homogeneous plane waves. However, the plane wave spectrum of the line source contains both homogeneous and inhomogeneous waves, and it is the latter which lead to the surface wave in Eqn. (14).[15]

In general, the fields H_z' and H_z'' will exhibit exponential attenuation in both the vertical and horizontal directions. The horizontal attenuation length ℓ_H is the same for either field and is given by:

$$\ell_H = \frac{1}{k_o B} \tag{15}$$

$$B = -\mathrm{Im}\sqrt{\epsilon/1 + \epsilon}$$

When the medium is lossless ($\epsilon_2 = 0$), ℓ_H becomes infinite and there is no horizontal attenuation. For the field above the surface, H_z', the vertical attenuation length ℓ_v' is:

$$\ell_v' = \frac{1}{k_o D_2} \tag{16}$$

$$D_2 = \mathrm{Im}(1/\sqrt{1 + \epsilon})$$

and below the surface the attenuation length is:

$$\ell_v'' = \frac{1}{k_o G} \tag{17}$$

$$G = -\mathrm{Im}(\epsilon/\sqrt{1 + \epsilon})$$

However, even for a lossless material, both ℓ_v' and ℓ_v'' remain finite, i.e. the wave is still bound to the surface.

The phase velocity of the surface wave in the x direction is the same on either side of the interface and is given by:

$$v_p^x = \frac{\omega}{k_x} = \frac{c}{A} \tag{18}$$

$$A = \mathrm{Re}\sqrt{\frac{\epsilon}{1 + \epsilon}}$$

Since $A > 1$, the phase velocity is less than c and the surface waves are slow waves.

The time averaged power flux $< \bar{S} >$ is easily calculated from the following equation:

$$< \bar{S} > = \tfrac{1}{2}\,\mathrm{Re}\,[\bar{E} \times \bar{H}^*] \tag{19}$$

For the field above the surface ($x > 0$, $y > 0$), we find:

$$< \bar{S} >' = \tfrac{1}{2}\sqrt{\frac{\mu_o}{\epsilon_o}}\ \mathrm{Re}\left[\sqrt{\frac{\epsilon}{1+\epsilon}},\ -\frac{1}{\sqrt{1+\epsilon}},\ 0\right] \left| H_z' \right|^2$$

(20)

$$\left| H_z' \right|^2 = 8\pi \left| \frac{\epsilon^{3/2}}{\epsilon^2 - 1} \right|^2 \exp\left(\frac{-2x}{\ell_H} - \frac{2(y+h)}{\ell_v'}\right)$$

Below the interface ($x > 0$, $y < 0$) the power flux is:

$$< \bar{S} >'' = \tfrac{1}{2}\sqrt{\frac{\mu_o}{\epsilon_o}}\ \mathrm{Re}\left[\frac{1}{\sqrt{\epsilon}\sqrt{1+\epsilon}},\ -\frac{1}{\sqrt{1+\epsilon}},\ 0\right] \left| H_z'' \right|^2$$

(21)

$$\left| H_z'' \right|^2 = 8\pi \left| \frac{\epsilon^{3/2}}{\epsilon^2 - 1} \right|^2 \exp\left(\frac{-2x}{\ell_H} + \frac{2y}{\ell_v''} - \frac{2h}{\ell_v'}\right)$$

Several comments need to be made concerning Eqns. (20) and (21). The first concerns the factor $\exp(-2h/\ell_v')$ which is present in both $< \bar{S} >'$ and $< \bar{S} >''$. This is the "height-loss" factor, well-known in antenna studies, and indicates that the power coupled into the surface wave is a maximum when the source is on the interface ($h = 0$). As the source is moved away from the surface, the power in the surface wave will decrease exponentially.[16] Secondly, the power flux $< \bar{S} >'$ is at an angle $-\sigma_\rho$ with respect to the positive x axis. In general, power flows into the surface and accounts for ohmic losses. For a lossles medium, $\sigma_\rho = 0$, and the power flow is parallel to the interface. Similarly the power flux $< \bar{S} >''$ is directed into the medium in the lossy case. However, when $\epsilon_2 = 0$, $< \bar{S} >''$ is parallel to the interface but opposite to $< \bar{S} >'$ [17].

Now we consider the power flux per unit length in the z direction across the plane $x = x_o$. This quantity is denoted by $P_x(x_o)$ and is defined as

$$P_x(x_o) = \int_o^\infty < \bar{S} >' \cdot \hat{x}\ dy + \int_{-\infty}^o < \bar{S} >'' \cdot \hat{x}\ dy \tag{22}$$

Using Eqns. (20) and (21) for the power flux, we obtain:

$$P_x(x_o) = 8\pi\sqrt{\frac{\mu_o}{\epsilon_o}} \left| \frac{\epsilon^{3/2}}{\epsilon^2 - 1} \right|^2 \exp\left(-\frac{2x_o}{\ell_H}\right) \cdot$$

(23)

$$\exp\left(-\frac{2h}{\ell_v'}\right)\left[\ell_v' \ \mathrm{Re}\sqrt{\frac{\epsilon}{1+\epsilon}} + \ell_v''\ \mathrm{Re}\left[\frac{1}{\sqrt{\epsilon + \epsilon^2}}\right]\right]$$

Then the power coupled into the surface wave P_s is given by:[18]

$$P_s = 2\, P_x(0)$$

$$= \left\{ 8\pi \sqrt{\frac{\mu_o}{\epsilon_o}} \left| \frac{\epsilon^{3/2}}{\epsilon^2 - 1} \right|^2 \exp\left(- \frac{2h}{\ell_v'} \right) \right.$$

(24)

$$\left. \left[\ell_v' \, \mathrm{Re} \sqrt{\frac{\epsilon}{1+\epsilon}} + \ell_v'' \, \mathrm{Re} \left[\frac{1}{\sqrt{\epsilon}} \frac{1}{\sqrt{1+\epsilon}} \right] \right] \right\}$$

For large values of $|\epsilon|$, the power carried off by the surface wave is predominantly due to the field above the interface and is proportional to $|\epsilon|^{-\frac{3}{2}}$. Also, from the factor, $\exp(-2h/\ell_v')$, this power will be appreciable only if $h \lesssim \ell_v'$, i.e., the line source must be located within the vertical attenuation length of the resulting field.

5. CONCLUDING REMARKS. We have shown that with a p-polarized source it is possible to excite surface waves directly on a homogeneous, non-magnetic half-space provided that the dielectric response of the medium is such that $\mathrm{Re}\sqrt{\epsilon/(1+\epsilon)} > 1$. For metals, this implies that the frequency of the source must be below the plasma frequency of the metal. The resulting surface waves are Zenneck waves and exhibit exponential attenuation away from the interface. We have also derived explicit expressions for the attenuation lengths and for the Poynting vector of these fields.

The duality principle implies that similarly, it is possible to excite surface waves on the planar interface of media having different magnetic permeabilities using an s-polarized source. Our results can be restated to apply to this case by simply substituting $E \to H$, $H \to -E$, and $\mu \Leftrightarrow \epsilon$ in the above formulae.

The present investigation represents one facet of our continuing study of the electromagnetic (optical, infrared) properties of surfaces. These results form an intermediate step in the calculation of the launching efficiency of electromagnetic surface waves on a corrugated or randomly rough surface. Such calculations would apply to the coupling of electromagnetic radiation into surface structures and relates to problems connected with laser mirrors, surface-wave devices, and integrated optics.

REFERENCES

1. For more information concerning surface plasmons see Polaritons, E. Burstein and F. DeMartini, Eds., (Pergamon, New York, 1974).

2. A. Otto, Phys. Stat. Sol. 26, K99 (1968).

3. J.D. McMullen, Technical Report, U.S. Army Research Office, 1974.

4. A number of such devices are described by F.A. Blum,Microwaves, 14 No. 5, 56 (1975).

5. E.L. Church and J.M. Zavada, J. Opt. Soc. Am. 64, 547A (1974).

6. E.L. Church, J.M. Zavada and H.N. Kritikos, Bull. Am. Soc. 588 (1975).

7. Our original study was in connection with metal surfaces. However, this calculation is also valid for the interaction of EM waves with a plasma medium. In the course of preparing this paper, we have come across a similar problem treated in L. Felsen and N. Marcuvitz, Radiation and Scattering of Waves (Prentice-Hall, Englewood Cliffs, 1973).

8. P.C. Clemmow, The Plane Wave Spectrum Representation of Electromagnetic Fields, (Pergamon, New York, 1966).

9. For a discussion of inhomogeneous waves see J.A. Stratton, Electromagnetic Theory, (McGraw-Hill, New York, 1941).

10. Saddle point integration is discussed in many places such as: Felsen and Marcuvitz, op. cit.; R.E. Collin, Field Theory of Guided Waves, (McGraw-Hill, New York, 1960); and G. Tyras, Radiation and Propagation of Electromagnetic Waves, (Academic, New York, 1969).

11. The properties of lateral waves are discussed in Tyras, ibid.

12. If we had used an s-polarized line source, the reflected field would be similar to the integral in Eqn. (8) except that $\rho_{\parallel}(\alpha)$ would be replaced by $\rho_E(\alpha)$, the Fresnel coefficient for s-polarization. For non-magnetic materials $\rho_E(\alpha)$ is free of poles and, consequently, no surface waves are generated in this problem with an s-polarized source.

13. Collin, op. cit., pp. 501, 502.

14. Zenneck Waves are treated at length in H.E.M. Barlow and J. Brown, Radio Surface Waves, (Oxford, London, 1962).

15. The importance of the inhomogeneous waves in the spectrum of the source is brought out in A.L. Cullen, Proc. IEEE 101, Part IV 255, (1954).

16. Cullen, ibid.

17. This unusual behavior of the Poynting vector is discussed in J.S. Schoenwald's thesis, Univ. of Penna., 1973.

18. The problem is symmetric with respect to the y axis. Two surface waves are excited and, hence, the factor 2 in Eqn. (24).

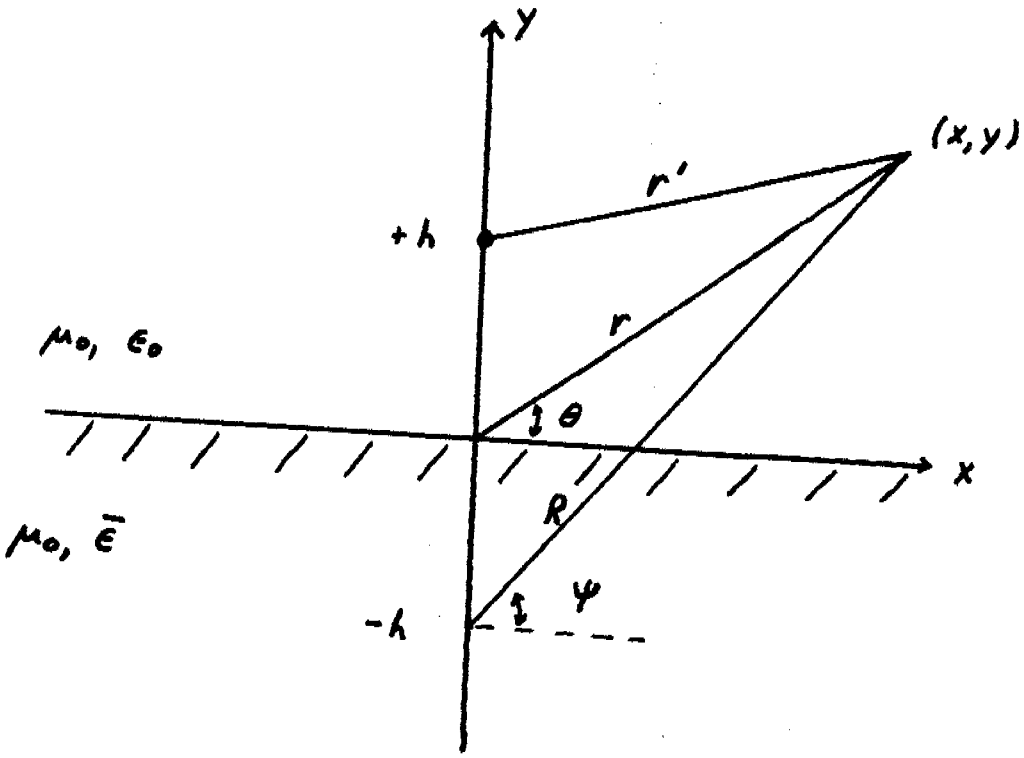


FIGURE 1

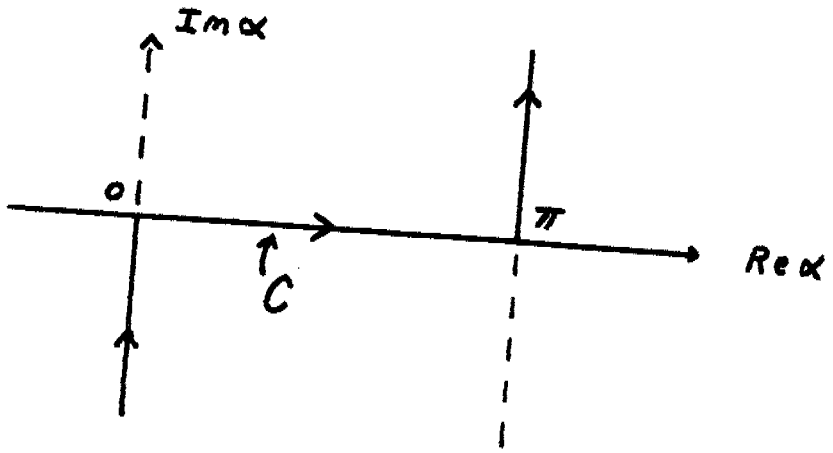


FIGURE 2

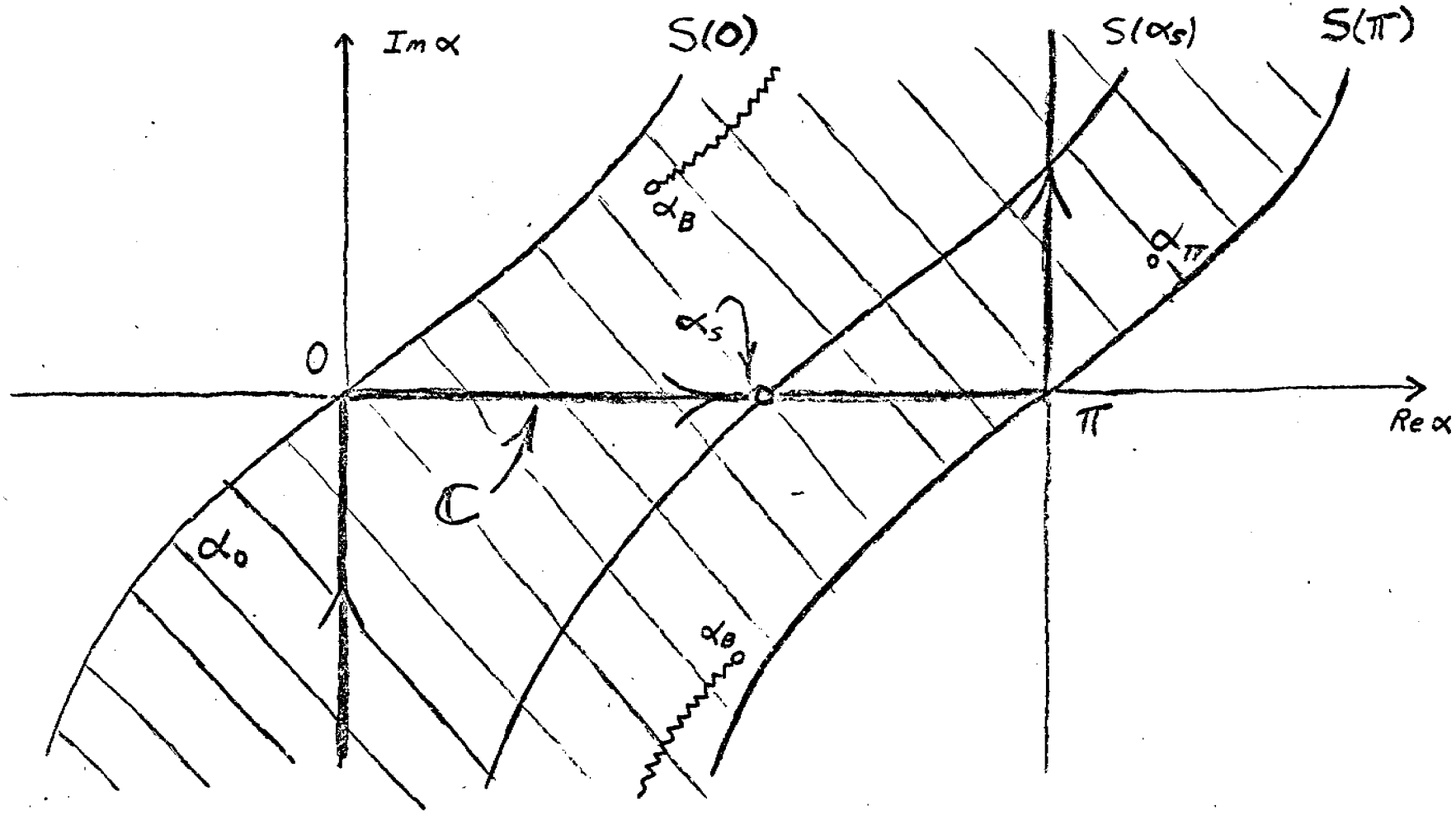


FIGURE 3

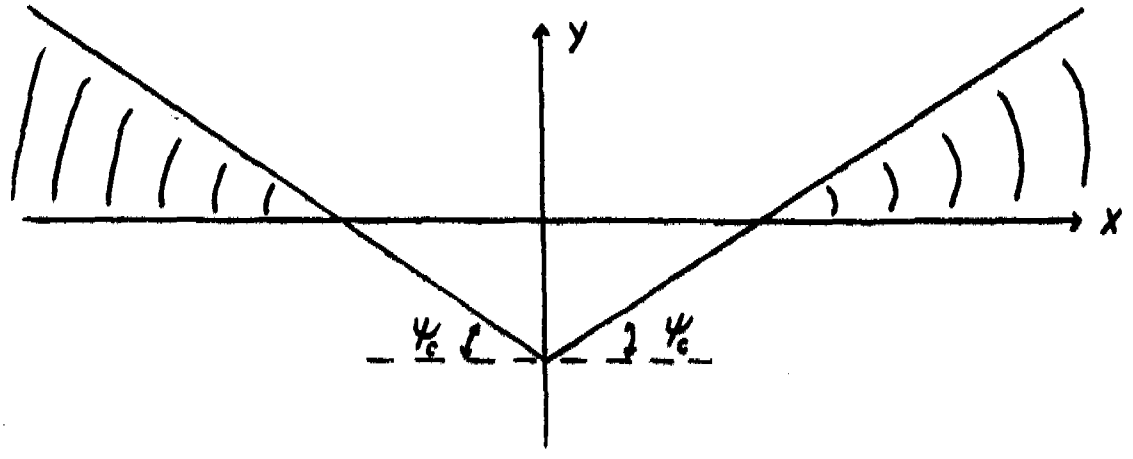


FIGURE 4

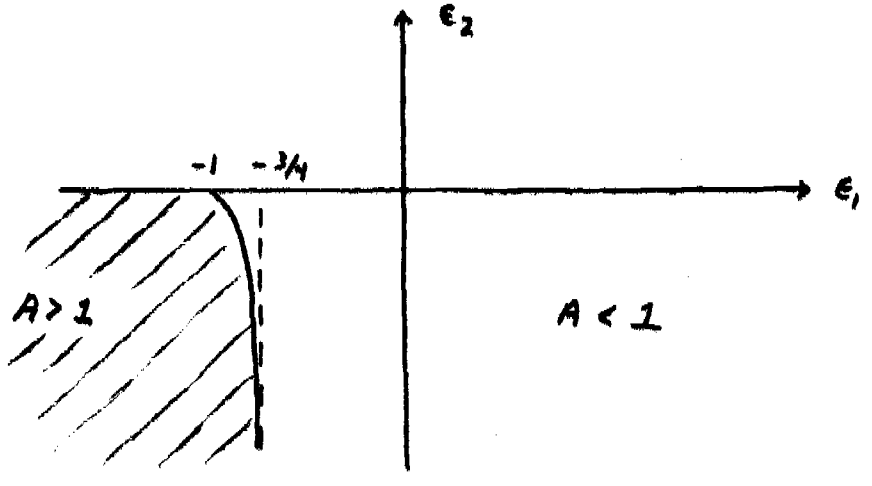


FIGURE 5

COMPARISON OF PERTURBATION-THEORETIC AND EXACT CALCULATIONS OF NONLINEAR OPTICAL PROPERTIES OF OPTOELECTRONIC MATERIALS

S. S. Mitra,* L. M. Narducci,† and R. A. Shatas
Quantum Physics, Physical Sciences Directorate, US Army Missile Command
Redstone Arsenal, AL 35809

ABSTRACT. Nonlinear absorption coefficients due to multiphoton transitions induced by intense radiation fields have been calculated for direct bandgap semiconductors at 0.964, 1.06, 1.318, and 10.6 μm laser wavelengths and compared with experimental results. Second-order perturbation models of Braunstein (interconduction band transitions) and Basov (intravalence or intraconduction band transitions) in their original form yield underestimates and slight overestimates of the nonlinear absorption, respectively. Corrections to these models based on a detailed consideration of effective band mass and interaction parameters are presented which decrease the discrepancy between the theory and the experiment. The Keldysh model gives second-order absorbances that are in between both perturbation calculations; also, it predicts very well the one-photon band-edge absorbance in GaAs and InSb. All models predict only the correct order of nonlinear absorption coefficient (typical $\alpha_2 = 10^{-6}...10^{-8}$ cm W^{-1}). Inclusion of exciton bands is needed for a close agreement between theory and experiment. This suggests that the incorporation of details of energy bands in solids in semiclassical nonlinear optoelectronic calculations is more important than a full accounting of quantum electrodynamical corrections.

1. INTRODUCTION. Under illumination by intense coherent infrared radiation sources presently available, many optoelectronic materials are driven into a nonlinear response region. In this paper, we analyze processes contributing to the nonlinear absorption of light by electronic transitions taking place as a result of a multiquantal process in a dielectric or a semiconductor material usually transparent to low intensity radiation. We also show that the absorption at high radiation fluxes is describable by a nonlinear differential equation. For the case of a weak radiation field, this equation can be linearized and its solution yields the Lambert-Beer law. At high intensities, however, nonlinear terms dominate, and consequently, a general analytic solution is not available. Physical arguments must be invoked to separate terms in powers of intensity. For each particular ratio of materials properties to radiation flux parameters, one particular term dominates the absorption. Electronic transition rates induced by photon processes of appropriate multiplicity enter as coefficients in the nonlinear differential equation describing the absorption. Their calculation can be undertaken in the framework of the time-dependent perturbation theory of wave mechanics; also, a semiphenomenological approach employing the rudiments of the "dressed state" concept of the

*Permanent address: Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881.
†Permanent address: Department of Physics, Worcester Polytechnic Institute, Worcester, MA 01609.

quantum field theory is parametrized in terms of the one-electron effective mass approximation. In the perturbation-theoretic approach, a semiclassical treatment of the interaction operator is undertaken in which the vector potential of the Maxwell field is used to describe the optical radiation, and the oscillating electron in the momentum representation accounts for the electronic motion. The description is complicated by the need to employ the effective mass parametrization of the energy band theory of the solid state. Within the limits imposed by these fundamental difficulties, numerical methods were developed to calculate multiphoton absorption for a number of technologically important semiconductors. Because of a rather unsatisfactory state of the fundamental knowledge in the nonlinear quantum optics from both the physical and the mathematical points of view, a deeper insight through the employ of methods leaning towards the axiomatic quantum field theory was sought. One significant aspect of this approach is that the dynamic eigenlevel shifts of the originally quasi-stationary states caused by the turn-on of the radiation field can be incorporated at the beginning of calculations. This fast-varying modulation of the eigenfunction of a discrete state of the matter field is not usually accessible by the perturbation expansion of the first-quantized wave mechanics simply because the order of perturbation expansion usually is not carried out beyond the first-nonzero contribution. Computerized numerical solutions for transition probabilities in various approaches are given and compared with experimental data; disagreements between theory and experiment are shown to arise from certain inadequacies of the theoretical approach. Suggestions with respect to the further development of the nonlinear quantum optics of solids are offered. In reviewing the research literature on the multiphoton absorption processes in gases, we note extensive investigations in recent years.[1-3] Unfortunately, despite numerous theoretical and experimental contributions, there is a wide scatter between both the measured and the predicted values of multiphoton absorption. Because of the many applications of pulsed lasers, there is a need to identify theoretical approaches which are suitable to predict at least the order of magnitude of multiphoton absorption in optically transparent materials. In the case of gases, the ratio of the ionization potential to the photon energy of a typical pulsed laser is rather high, therefore atomic photoionization experiments are not suitable to test theories constructed for optical electromagnetic fields of moderate intensity. Specifically in what follows, we compare two low-order perturbation treatments with a semiclassical procedure in which the band distortion due to the electromagnetic field is taken into account. Within the context of models proposed by Keldysh,[4] Braunstein and Ockman,[5] and Basov et al.,[6,7] we calculate second-order absorption coefficients for a number of direct-bandgap semiconductors and compare theoretical predictions with available experimental results. In section 2, we consider the rate equations describing the nonlinear absorption. In section 3, we outline a general formulation of the multiphoton absorption within the semiclassical radiation theory in the electric dipole approximation. Section 4 deals with the Keldysh "exact" model which employs conduction and valence band electronic wavefunction "dressed" by the radiation field. Section 5 contains the perturbation-theoretic approaches of Braunstein and Ockman. Comparison between the theoretically predicted and the experimentally determined values of nonlinear absorption coefficients is given in section 6.

2. ATTENUATION OF LIGHT BY NONLINEAR ABSORPTION. In a transparent material, residual absorption of light may be related to a number of independent processes such as multiphoton and -phonon excitations, presence of residual free carriers or impurities, creation of excitonic states, and perhaps phonon parametric amplification. At sufficiently high intensities of the optical field, free carriers created through the elementary excitation processes enumerated above, may contribute to the time dependence of the absorption through the free carrier interaction with the electromagnetic field (so-called inverse Bremsstrahlung). The time dependence of this part of absorption follows from the rate equations describing the free carrier density and shows typically transients of 10^{-12} to 10^{-8} sec duration reflecting the trapping and recombination rates in solid materials. In these considerations, we will neglect these transient contributions.

For the steady-state description of the nonlinear absorption, we argue that for a given material with a fixed bandgap between the valence and conduction bands E_g and a given incident photon energy $\hbar\omega$, the photon absorption on ν-th order dominates. Here ν denotes the ratio $\langle E_g/\hbar\omega + 1\rangle$, and the brackets $\langle\ \rangle$ stand for the integer part of this ratio. Although linear energy losses caused by nonresonant processes will always be present even in the purest materials, low-intensity measurements can yield reliable values of the linear absorption coefficient. Hence observed intensity loss due to multiphoton absorption should be satisfactorily described by the phenomenological rate equation

$$\frac{dI}{d\ell} = -(\alpha_1 I^1 + \alpha_2 I^2 + \ldots + \alpha_n I^n) = -\sum_{\nu=1} \alpha_\nu I^\nu \qquad (2.1)$$

where ℓ represents the spatial coordinate along the direction of travel of the light in the material, and α_ν is the ν-photon absorption coefficient. The total experimentally observed intensity attenuation should be corrected for the contribution due to nonresonant losses. The corrected value provides the rate of intensity loss $dI/d\ell$ due to nonlinear absorption processes.

In the above rate equation, I denotes the incident intensity (light flux in units of W cm^{-2}). The dimension of the ν-th order absorption coefficient α_ν is $(\text{length})^{2\nu-3}/(\text{power})^{\nu-1}$. If the multiphoton process of order ν dominates, we can express the attenuation rate of the light flux as

$$\frac{dI}{dt} = -\frac{c}{\sqrt{\epsilon_\infty}} \sum_{\nu} \alpha_\nu I^\nu$$

where c is the velocity of light in vacuum, and ϵ_∞ is the high-frequency dielectric constant of the material. The relation between the flux in the material I [W cm^{-2}] and the peak electric field amplitude E_o [V cm^{-1}] in practical electromagnetic units is given by $I = \frac{1}{2}(E_o^2 \sqrt{\epsilon_\infty}/R_o)$, where R_o is the vacuum impedance (377 Ω). The relation between the flux

I [W cm^{-2}] and the photon number density N_{ph} [cm^{-3}] is given by $I = (c/\sqrt{\epsilon_\infty})\,\hbar\omega\, N_{ph}$. Accordingly, the rate of photon absorption per unit volume is expressed by

$$\frac{dN_{ph}}{dt} = -\frac{\alpha_\nu}{\hbar\omega} I^\nu = -\nu \frac{dN_e}{dt} \quad ,$$

where N_e designates the number density of free carriers created by across the gap ionization through the ν-photon absorption. The right-hand side of the above equation is also a consequence of the energy balance condition since the rate of the photon absorption must be ν-times larger than the rate of the free carrier creation; hence $dN_{ph}/dt + \nu(dN_e/dt) = 0$. We also designate

$$\frac{dN_e}{dt} = w(E_o^{2\nu}) \quad ,$$

where w is the multiphoton transition rate; consequently, the multiphoton absorption coefficient for the ν-th order process is given by

$$\alpha_\nu = \frac{2\nu\,\hbar\omega\,(2R_o)^\nu\; w(E_o^{2\nu})}{\epsilon_\infty^{\nu/2}\; E_o^{2\nu}} \quad . \tag{2.2}$$

In eq. (2.2) ϵ_∞ is the high frequency dielectric constant, and the factor of 2 is introduced to account for the free carrier spin degeneracy. Note that despite the apparent dependence of the absorption coefficient upon E_o, the actual calculated numerical values of α_ν given in section 6 are field independent.

3. MULTIPHOTON ABSORPTION BY THE SEMICLASSICAL TIME-DEPENDENT PERTURBATION. The theoretical description of multiphoton absorption is based on the employ of second- and higher-order time-dependent quantum mechanical perturbation theory. In the case of the two-photon process, it has been first shown by M. Goeppert-Mayer[8] that the interaction term coupling the momentum of the electron $\vec{p}$ with the vector potential of the Maxwell field $\vec{A}$, given by $(e/mc)\,\vec{p}\cdot\vec{A}$, can be canonically transformed to $\vec{P}\cdot\vec{E}$. This transformation is valid when the linear extent of the interacting electron-ion system is small with respect to the wavelength of the radiation field (this condition is readily satisfied for the electric dipole mediated transitions into far ultraviolet wavelengths); subsequently, either of these interaction terms can be used in perturbation expansion.

At first we shall review the salient features of the perturbation theory. Let there be two eigenstates of energy E_i and E_f; we are concerned with an electric dipole transition between them induced by photons of energy $\hbar\omega < E_f - E_i$. Such transitions can then only take place by the participation of more than one photon. The order of the photon process is obviously given by $\langle 1 + (E_f - E_i)/\hbar\omega \rangle$ where $\langle\ \rangle$ indicates the integer part. The probability of such transitions will depend

on the number density of photons or the electric field intensity of the radiation. Large field intensities will also introduce level shifts.

Let us consider the time-dependent Schrödinger equation

$$\mathcal{H}\psi = i\hbar \frac{\partial \psi}{\partial t} \quad . \tag{3.1}$$

The Hamiltonian $\mathcal{H}$ consists of the unperturbed time-independent Hamiltonian $\mathcal{H}^o$ and a time-dependent perturbation term $\mathcal{H}'(\vec{r},t)$. If $\psi_n^o(\vec{r})$ defined by

$$\mathcal{H}^o\psi_n^o(\vec{r}) - E_n\psi_n^o(\vec{r}) = 0 \tag{3.2}$$

are the eigenstates of $\mathcal{H}^o$, $\psi_n^o(\vec{r},t)$ is given by

$$\psi_n^o(\vec{r},t) = \psi_n^o(\vec{r}) \exp\left[-i \frac{E_n}{\hbar} t\right] \quad . \tag{3.3}$$

The general state function, which is the solution of eq. (3.1), can be expressed as a superposition of states

$$\psi(\vec{r},t) = \sum_n a_n(t)\, \psi_n^o(\vec{r},t) \quad . \tag{3.4}$$

The probability that at any time t the system is in a state with energy E_n is given by $|a_n(t)|^2$. Substituting eq. (3.4) in eq. (3.1), and making use of eqs. (3.2) and (3.3), one obtains

$$\dot{a}_n(t) = \frac{1}{i\hbar} \sum_{n'} a_{n'}(t) \langle \psi_n | \mathcal{H}' | \psi_{n'} \rangle \quad . \tag{3.5}$$

Equation (3.5) may be used to solve a_n to any order. The zeroth order coefficients $a_n^{(o)}(0)$ at the time $t = 0$ are zero except for the initial state, for which

$$a_i^{(o)}(0) = 1 \quad .$$

The first order coefficient $a_n^{(1)}(t)$ is therefore

$$a_n^{(1)}(t) = \frac{1}{i\hbar} \int_o^t \langle \psi_n | \mathcal{H}' | \psi_i \rangle \, dt' \quad . \tag{3.6}$$

To obtain the second order correction to the coefficient $a_n(t)$, one

substitutes the above value of $a_n^{(1)}(t)$ in the right-hand side of eq. (3.1) and solves for $a_f^{(2)}(t)$. Continuing this iterative process, one can write for a general n-th-order correction

$$a_f^{(n)}(t) = \left(\frac{1}{i\hbar}\right)^n \int_0^t \langle \psi_f | \mathcal{H}' | \psi_n \rangle \, dt' \int_0^{t'} \langle \psi_n | \mathcal{H}' | \psi_m \rangle \, dt'' \ldots$$

$$\times \int_0^{t^{n-1}} \langle \psi_s | \mathcal{H}' | \psi_i \rangle \, dt^{(n)} \quad . \tag{3.7}$$

The repeated indices of the wave functions are the so-called intermediate states. With the use of eq. (3.3), eq. (3.7) may be rewritten in terms of the unperturbed wave functions as

$$a_f^{(n)}(t) = \left(\frac{1}{i\hbar}\right)^n \left[\frac{\langle \psi_f^o | \mathcal{H}' | \psi_n^o \rangle}{E_n - E_r \pm \hbar\omega_1 \ldots \pm \hbar\omega_n} \cdots \right.$$

$$\left. \times \frac{\langle \psi_s^o | \mathcal{H}' | \psi_i^o \rangle}{E_s - E_i \pm \hbar\omega_1} \right] \int_0^t \exp\left\{ \frac{i(E_f - E_i \pm \hbar\omega_1 \ldots \pm \hbar\omega_n)t'}{\hbar} \right\} dt' \quad . \tag{3.8}$$

For the sake of generality, every one of the n-photons has been assigned a different frequency $\omega_n \neq \omega_m$.

The transition probability rate may be readily obtained from (3.8) and is given by

$$w_{if}(\omega) = \frac{d}{dt} \left| a_{if}(t) \right|^2 \quad . \tag{3.9}$$

So w_{if} will contain square of the expression in the square bracket of eq. (3.8), and the integral in eq. (3.8) will introduce a δ-function

$$\delta(E_f - E_i \pm \hbar\omega_1 \ldots \pm \hbar\omega_n)$$

to satisfy the energy conservation requirement.

For optical transitions in a crystalline solid, the initial and final states are obviously the valence and the conduction bands, respectively. The choice of the intermediate states accounts for different results in different calculational schemes. The order of the transition equals the number of quanta of the perturbing field absorbed or emitted. The intermediate states do not appear in the δ-function but in the denominator of the square of the pre-integral terms of eq. (3.8). Therefore the contribution of the intermediate states to the transition probability

becomes significant when they are located close to the initial or final states.

In the presence of a radiation field described by the vector potential $\vec{A}$, the kinetic energy of the electron is modified to

$$\frac{1}{2m}\left(\vec{p} - \frac{e}{c}\vec{A}\right)^2$$

which as a result introduces a perturbation Hamiltonian in eq. (3.1) of the type

$$\mathcal{H}' = \frac{1}{mc}\vec{p} \cdot \vec{A} \tag{3.10}$$

where $\vec{A} = A_o \hat{e}_1 \exp[i\vec{k} \cdot \vec{r} - \omega t]$. Here, $\vec{k}$ and ω are photon wave vector and frequency, respectively, and $\hat{e}$ its unit polarization vector. In eq. (3.10) we have neglected the $|\vec{A}|^2$ term, which is usually very small at the wavelengths and intensities in question. The amplitude $\hat{e}A_o$ of the vector potential is related to the photon density N_{ph} by

$$A_o^2 = \frac{2\pi N_{ph} \hbar c^2}{\omega \epsilon_\infty} \text{ (c.g.s.)} \tag{3.11}$$

where ϵ_∞ is the dielectric constant. The calculation of the transition probability rate then boils down to the calculation of the matrix elements of $\vec{p} \cdot \vec{A}$ between the valence band (or bands), the chosen intermediate levels and the conduction band. This is essentially what Braunstein et al.[5] and Basov et al.[6,7] models pertain to do.

Keldysh treatment,[4] however, differs from the above scheme as it also includes the level shifts caused by the perturbing electric field $\vec{E}$, which was neglected by Braunstein and Basov. The energy shift of a level induced by the field $\vec{E} = \hat{e}\vec{E}_o \exp[i(k \cdot \vec{r} - \omega t)]$ is given by

$$\Delta E_n = \frac{1}{4}\sum_m \left\{ \frac{|\vec{p}_{mn} \cdot \vec{E}_o|^2}{E_n - E_m - \hbar\omega} + \frac{|\vec{p}_{mn} \cdot \vec{E}_o|^2}{E_n - E_m + \hbar\omega} \right\} . \tag{3.12}$$

where the summation extends to all other states of the system. In the case of a semiconductor, these level shifts account for the change in bandgap in the presence of a field, which in the case of static fields is fairly well known, and is termed the Franz-Keldysh effect.[9] Keldysh's treatment for the multiphoton transition probability between shifted levels incorporates the dynamic Franz-Keldysh effect, and will be described in section 4.

4. THE KELDYSH "DRESSED STATE" MODEL. Keldysh[4] treated the multiphoton absorption as the high frequency limit of the time-dependent tunneling induced by the oscillatory electric field of the laser radiation. In other words, a unified description is available for the autoionization process under the influence of a strong low-frequency field and for the multiphoton ionization induced by a strong high-frequency field. Keldysh's treatment applies to both isolated atoms and crystalline solids (e.g., semiconductors). In both cases Keldysh calculates the probability of direct electron excitation from the ground state (or valence band) to the continuum (or conduction band) and the excitation probability through intermediate states of the discrete spectrum. The intermediate states could be higher excited states of the Coulomb field of isolated atoms, or impurity levels and excitonic states in crystalline solids.

Keldysh considers a system described by a Hamiltonian of the form

$$\mathcal{H} = \mathcal{H}_o + \mathcal{H}_C + \mathcal{H}_F \tag{4.1}$$

where $\mathcal{H}_o$ is the kinetic energy term, $\mathcal{H}_C$ is the Coulomb contribution to the potential energy (or the crystal field in solids), and $\mathcal{H}_F$ is the interaction energy of the bound charge in a periodic electric field $E(t) = E_o \cos \omega t$. The electron is assumed to be initially in the unperturbed Coulomb ground state given by the wave function

$$\begin{aligned} \psi_o(r,t) &= \psi_o(r) \exp(-i\mathcal{E}_o t/\hbar) \\ &= \frac{1}{(\pi a_o^3)^{\frac{1}{2}}} \exp(-r/a_o) \exp(-i\mathcal{E}_o t/\hbar) \end{aligned} \tag{4.2}$$

where $a_o = \hbar^2/(me^2z)$ is the z-charged atom Bohr radius and $\mathcal{E}_o$ its ionization potential.

In the case of a crystalline solid, the Bloch wave function is modified by the presence of the electromagnetic field as follows. (In a sense, the bare electron state is "dressed" by the electric field component of the e.m. field.)

$$\psi_p^{(v)}(\vec{r},t) = u_{p(t)}^{(v)}(\vec{r}) \exp\left\{\frac{i}{\hbar}\left(\vec{p}(t)\cdot\vec{r} - \int_o^t d\tau E_v[p(\tau)]\right)\right\} \tag{4.3}$$

where

$$\vec{p}(t) = \vec{p} + \frac{e\vec{E}_o}{\omega}\sin\omega t \quad . \tag{4.4}$$

The amplitude functions $u_{p(t)}^{(v)}(\vec{r})$ correspond to a valence electron with

momentum p(t) and have the translational symmetry of the lattice. The principal difference between the traditional treatment of this problem and the Keldysh approach is that he recognizes the modification of the unperturbed bands under the influence of the field. This modification amounts to a replacement of the unperturbed momentum $\vec{p}$ with the time-dependent momentum given by eq. (4.4). The excited electronic states associated with the continuum of an atom or in the conduction band of a solid are assumed to be unaffected by the electrostatic Coulomb energy, but are modified by the perturbing effect of the electromagnetic field. Thus, in the case of isolated atoms, the ionized electron is described by the solution of the Schrödinger equation

$$i\hbar \frac{\partial \psi}{\partial t} = -\left(\frac{\hbar^2}{2m}\nabla^2 + e\vec{E}(t) \cdot \vec{r}\right)\psi \tag{4.5}$$

which is given by

$$\psi_p(\vec{r},t) = \exp\left\{\frac{i}{\hbar}\left(\vec{p} + \frac{e\vec{E}_o}{\omega}\sin\omega t\right)\cdot\vec{r}\right\}$$
$$\times \exp\left\{-\frac{i}{\hbar}\int_o^t d\tau \;\frac{1}{2m}\left(\vec{p} + \frac{e\vec{E}_o}{\omega}\sin\omega\tau\right)^2\right\} \quad . \tag{4.6}$$

For the case of a solid, the conduction electron is described by the Bloch function

$$\psi_p^{(c)}(\vec{r},t) = u_{p(t)}^{(c)}(\vec{r})\exp\left\{\frac{i}{\hbar}[\vec{p}(t)\cdot\vec{r} - \int_o^t d\tau E_c[p(\tau)]\right\} \quad . \tag{4.7}$$

One again notices that the wave function of the final state is also dressed by the perturbing action of the field. As a consequence, the electronic transition does not occur between unperturbed states, but rather between nonstationary states in which the electron acceleration due to the field is taken into account. Finally, the transition probability rate is calculated according to first order perturbation theory. The calculated transition probability is then summed over all possible final momenta of the quasi-free electron.

In order to arrive at an explicit expression for the transition probability, Keldysh used the following parabolic energy-band relation for a solid

$$E(k) = E_g\left(1 + \frac{\hbar^2 k^2}{m^* E_g}\right)^{\frac{1}{2}} \tag{4.8}$$

where m^* designates the reduced effective mass of the electron-hole pair $1/m^* = 1/m_e^* + 1/m_h^*$, E_g denotes the width of the bandgap, and the momentum and k-space representations are connected by the relation $\vec{p} = \hbar\vec{k}$.

In the limiting case when the parameter

$$\gamma = \frac{\omega}{eE_o} (2m^*E_g)^{\frac{1}{2}} \tag{4.9}$$

is much larger than unity (a condition which is readily satisfied for most of crystalline solids), Keldysh's transition rate (electronic transition probability per unit volume and per unit time) is given by

$$w = \frac{2}{9\pi} \omega \left(\frac{m^*\omega}{\hbar}\right)^{\frac{3}{2}} \Phi[(2\langle x + 1\rangle - 2x)^{\frac{1}{2}}]$$

$$\times \left(\frac{e^2E_o^2}{16m^*\omega^2E_g}\right)^{\langle x+1\rangle} \exp\left[2\langle x + 1\rangle \left(1 - \frac{e^2E_o^2}{4m^*\omega^2E_g}\right)\right] . \tag{4.10}$$

The meaning of symbols in eqs. (4.9) and (4.10) is given below: $\gamma = 6.45 \times 10^3 (m^*E_g)^{\frac{1}{2}}/(\lambda E_o)$, where m^* is given in units of electron mass, E_g in eV, λ-wavelength of incident light in vacuum (in μm), E_o is the electric field amplitude in the material in units of MV m^{-1}, $\Phi(z) = e^{-z^2} \int_o^z e^{y^2} dy$ is the Dawson integral, $x = E_g/\hbar\omega\left(1 + e^2E_o^2/(4m^*\omega^2E_g)\right)$, and $\langle \ldots \rangle$ is the integer part of the argument.

Up to five photon transition probability rates have been calculated by means of relation (4.10) for a number of semiconductors as function of wavelength of light at a given field intensity of 10 MV m^{-1} in the material. The relevant band masses and bandgaps used in these calculations are listed in Table 1.

Table 1. Material parameters used in calculation of multiphoton transition rates as function of energy of incident laser light shown in figs. 1 to 6.

Semiconductor	Bandgap (eV)	Effective mass ratio		
		Electron	Hole	Reduced Pair mass
GaAs	1.53	0.068	0.5	0.06
GaSb	0.8	0.047	0.5	0.043
InAs	0.46	0.02	0.41	0.019
PbS	0.34	0.66	0.5	0.364
InSb	0.26	0.013	0.6	0.0127
PbTe	0.24	0.22	0.29	0.125

The log w vs photon energy plots are given in figs. 1 to 6. Numbers written above each plateau of the transition rate indicate the integer

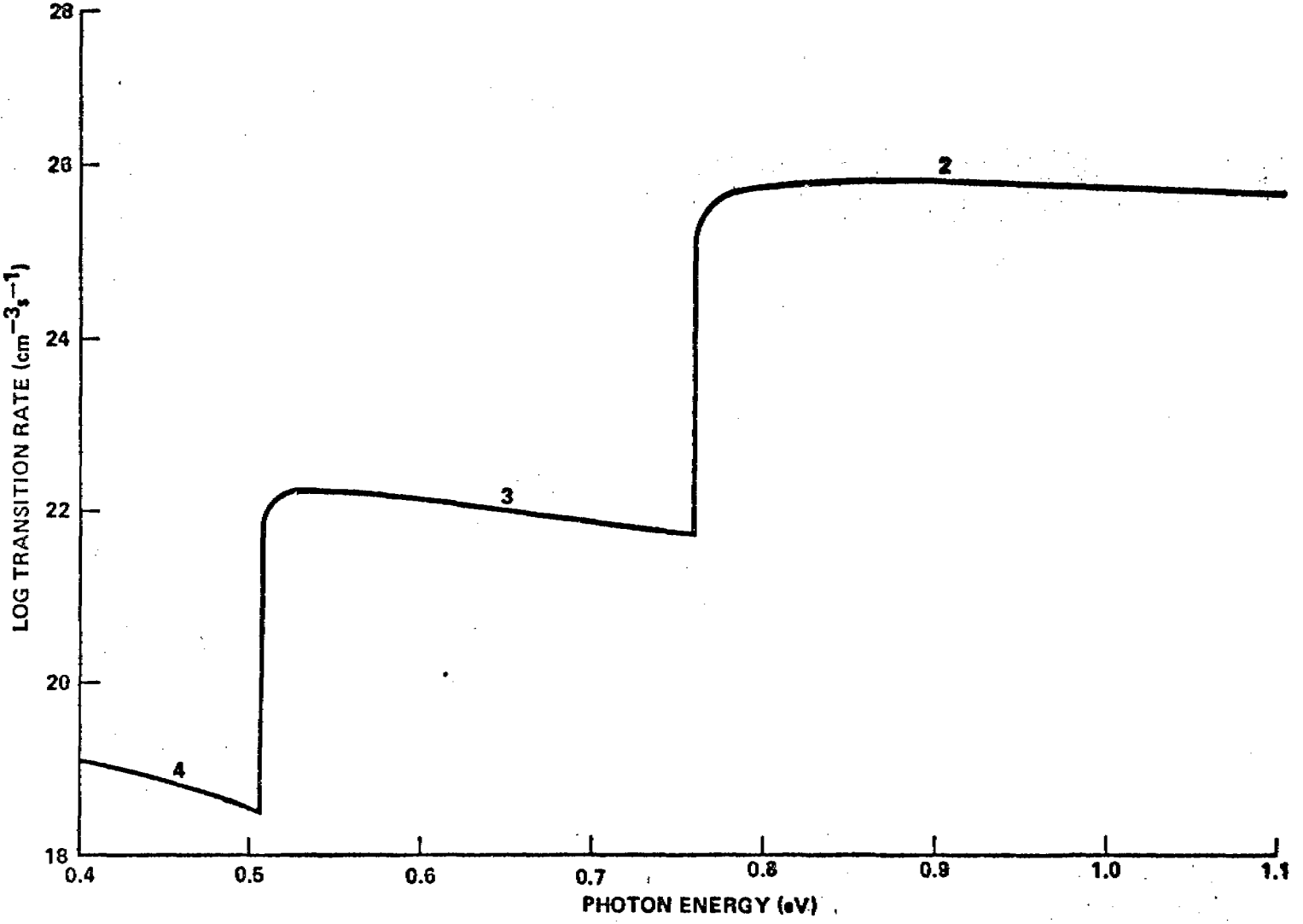


Fig. 1. GaAs multiphoton transition rate (2-nd to 4-th order) in cm^{-3} s^{-1} induced by a constant optical electric field of 10^4 V cm^{-1} amplitude and varied frequency.

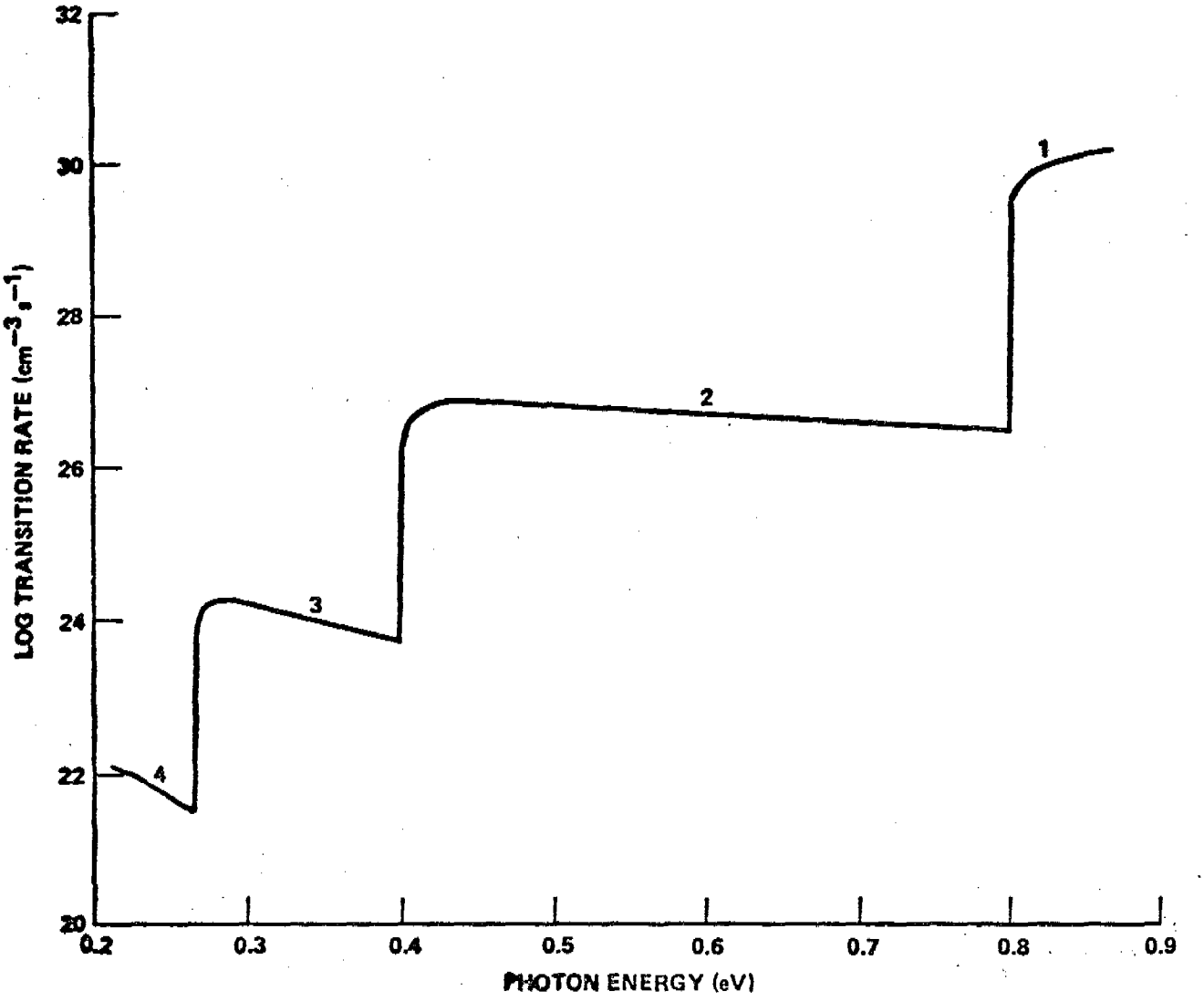


Fig. 2. GaSb multiphoton transition rate (1-st to 4-th order) in $cm^3\ s^{-1}$ induced by a constant optical electric field of 10^4 V cm^{-1} amplitude and varied frequency.

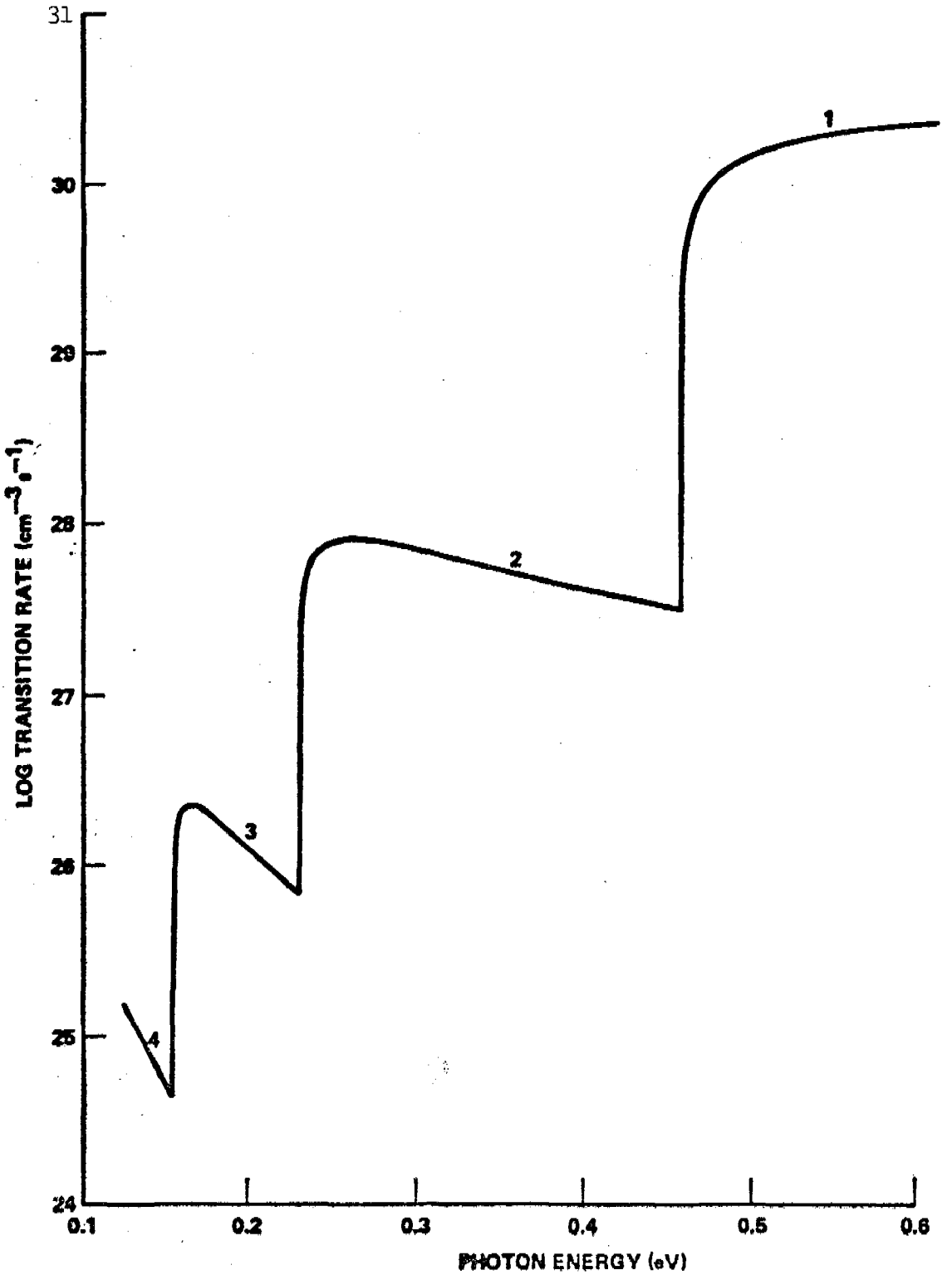


Fig. 3. InAs multiphoton transition rate (1-st to 4-th order) in cm^{-3} s^{-1} induced by a constant optical electric field of 10^4 V cm^{-1} amplitude and varied frequency.

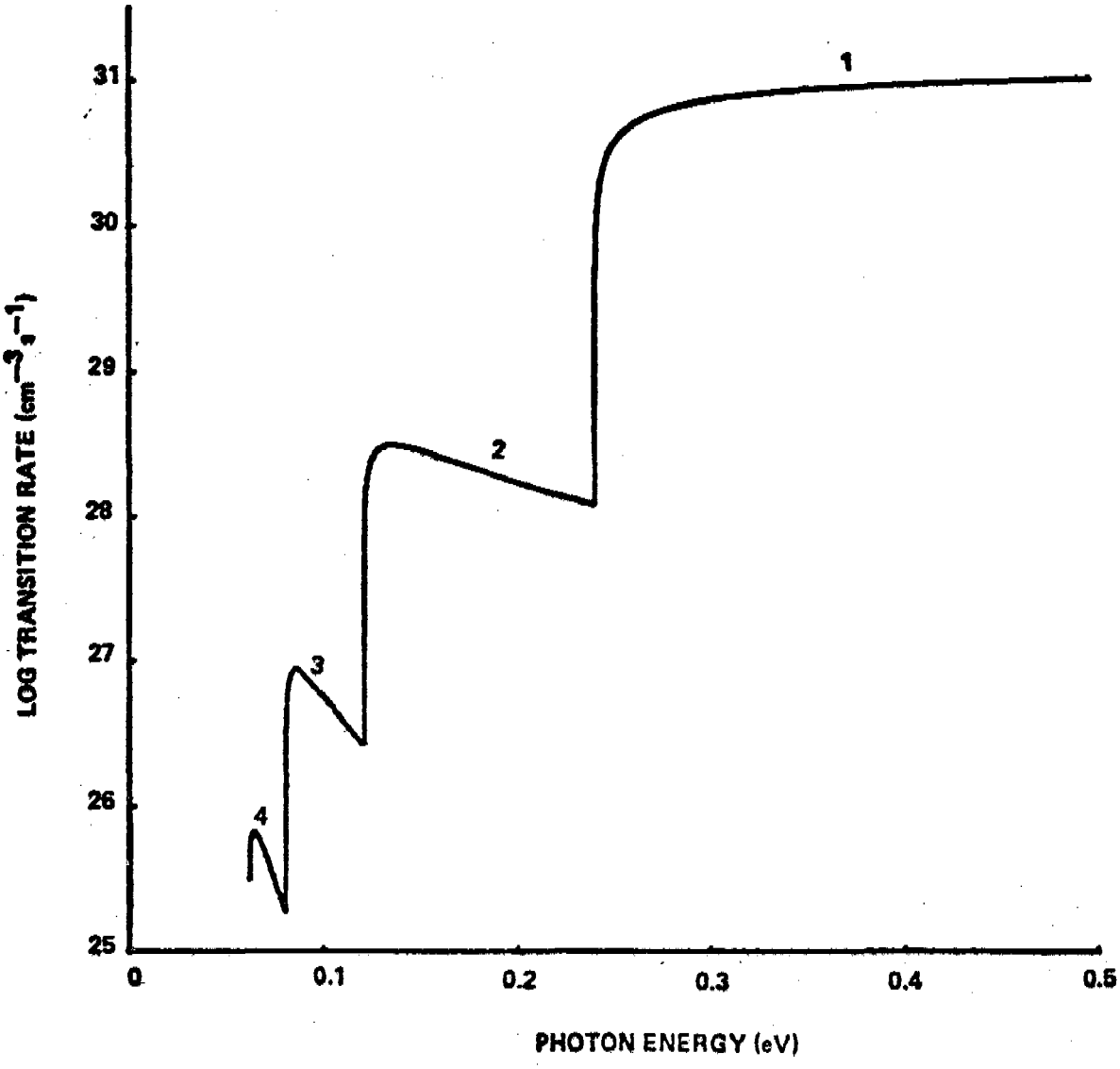


Fig. 6. PbTe multiphoton transition rate (1-st to 4-th order) in cm^{-3} s^{-1} induced by a constant optical electric field of 10^4 V cm^{-1} amplitude and varied frequency.

value of $\langle E_g/\hbar\omega + 1\rangle$. Apart from quantitative differences, the qualitative functional dependence in figs. 1 to 6 is strikingly similar in that a quasi-resonant behavior at integer $E_g/\hbar\omega$ values is displayed. This reminds us of one of the most obvious features of the multiphoton transition probability calculated by Bebb and Gold[10] for hydrogen-like atoms which shows the distinctly resonant dependence of the excitation probability as a function of the incident photon energy. While the energy levels of isolated atoms are clearly quite different from the bands of semiconductors considered in Keldysh's model, a quasi-resonant dependence of the transition rate on the incident photon energy seen in our calculations is qualitatively correct.

It should be also remarked that Weiler et al.[11] have extended Keldysh's calculation to include the effects of longitudinal or transverse magnetic fields on the interband electronic transition. Weiler's conclusions are in agreement with Keldysh's results for zero applied magnetic field. One expects a series of edge absorptions of the form $\sim (\nu\hbar\omega - E_g)^{\frac{1}{2}}$ where ν is the photon multiplicity. Such a behavior is indeed found in the negative slope of the transition rate after each quasi-resonant "peak" as seen in figs. 1 to 6. It has been argued that, at best, Keldysh's theory should provide acceptable results only for processes of fairly high photon multiplicity and that its application to two-photon absorption processes should not yield more than qualitative agreement with the experimental data. Recent experiments[12] involving four-photon transition in ZnS are at variance with Keldysh's prediction, while they are in fair agreement with perturbative semiclassical calculations.[13] However, a good agreement with three-photon absorption in CdS has been found. C. H. Lee measured[14] the three-photon absorption coefficient $\alpha_3 = 1.3 \cdot 10^{-2}$ cm^3/GW^2 for single-crystal CdS. Our calculations based on the evaluation of eqs. (2.2) and (4.10) yielded a value of $\alpha_3 = 0.2 \cdot 10^{-2}$ cm^3/GW^2. We have used $E_g = 2.42$ eV, $\epsilon_\infty = 5.32$ and $m^* = 0.192$ as CdS material parameters in this calculation. A more systematic comparison for the two-photon absorption coefficient is presented in section 6. An additional feature of the Keldysh model was found (which was rather unexpected) that it predicts very well the one-photon absorption coefficient near the band edge for two thus far analyzed semiconductors, GaAs and InSb. It is known that near the fundamental absorption edge, the one-photon absorption coefficient can be expressed as

$$\alpha_1 = A(\hbar\omega - E_g)^{\gamma} \tag{4.11}$$

where $\hbar\omega$ is the photon energy, and γ is a constant which equals 1/2 and 3/2 for allowed direct transitions and forbidden direct transitions, respectively. We specialize eq. (4.11) for the case $\hbar\omega > E_g$. In addition, the exponent γ equals 2 for indirect phonon assisted transitions and 1/2 for allowed indirect transitions to exciton states.

Near the absorption edge, where the values of ($\hbar\omega > E_g$) become comparable to the binding energy of the exciton, the Coulomb interaction between the free hole and the electron must be taken into account. For

$\hbar\omega < E_g$ the absorption merges continuously into that due to the higher excited states of the exciton; when $\hbar\omega \gg E_g$, higher energy bands will participate in the transition process and complicated band structures will be reflected in the absorption coefficient. The least square fit of absorbance vs. photon energy curves for photon energies near the band edge of GaAs yields the following values for A and γ : $A = 44 \cdot 10^3$, $\gamma = 0.499$, and $A = 44.7 \cdot 10^3$, $\gamma = 0.505$ from Moss and Hawkins'[15] and from Sturge's[16] measurements, respectively. Numerical evaluation of eqs. (2.2) and (4.10) for the one-photon transition case $\nu = 1$ can be expressed by the functional dependence shown in eq. (4.11) with $A = 44 \cdot 10^3$ and $\gamma = 1/2$. Hence, the Keldysh model besides giving the correct value of the absorption coefficient at the band edge of GaAs describes very well its wavelength dependence also. This agreement is seen in fig. 7 where unadjusted calculated absorption coefficients are compared with Moss and Hawkin's experimental values for photon energies between 1.42 to 1.48 eV. GaAs parameters used in calculating the theoretical absorption constant of fig. 7 are listed in Table 2.

Table 2. List of parameters for GaAs used in the theoretical fit of fig. 7. The values for reduced effective mass m* and dielectric constant ϵ_∞ have been obtained from [18] and [19], respectively.

Data source	E_g [eV]	m*	ϵ_∞
[15]	1.403	0.059	10.9
[16] (21^oK)	1.521	0.059	10.9
[16] (294^oK)	1.435	0.059	10.9

For the comparison of theoretical and experimental one-photon absorption coefficient of InSb, the experimental data points have been taken from fig. 3 of ref. 20. The empirical equation which fits the data around the band edge is listed in Table 3 together with the reduced effective mass and the high frequency dielectric constant which are needed for the theoretical comparison.

Table 3. Empirical relation for the absorption coefficients of InSb reported in [20]. The bandgap energy E_g has been calculated by a least square fitting procedure. The reduced effective mass m* and the high-frequency dielectric constant ϵ_∞ have been obtained from [19] and [21], respectively.

$$\alpha = 2.026 \cdot 10^4 \, (\hbar\omega - E_g)^{\frac{1}{2}} \, [cm^{-1}]$$

$\hbar\omega, E_g$ in eV

$E_g = 0.2248$ [eV]

$m^* = 0.0113\ m$; $m = 9.108 \cdot 10^{-28}$ [g]

$\epsilon_\infty = 15.7$

In fig. 8 we have plotted the experimental data together with the results of the numerical evaluation of eqs. (2.2) and (4.10). We note that the quantitative agreement between the Keldysh theory and the measured data

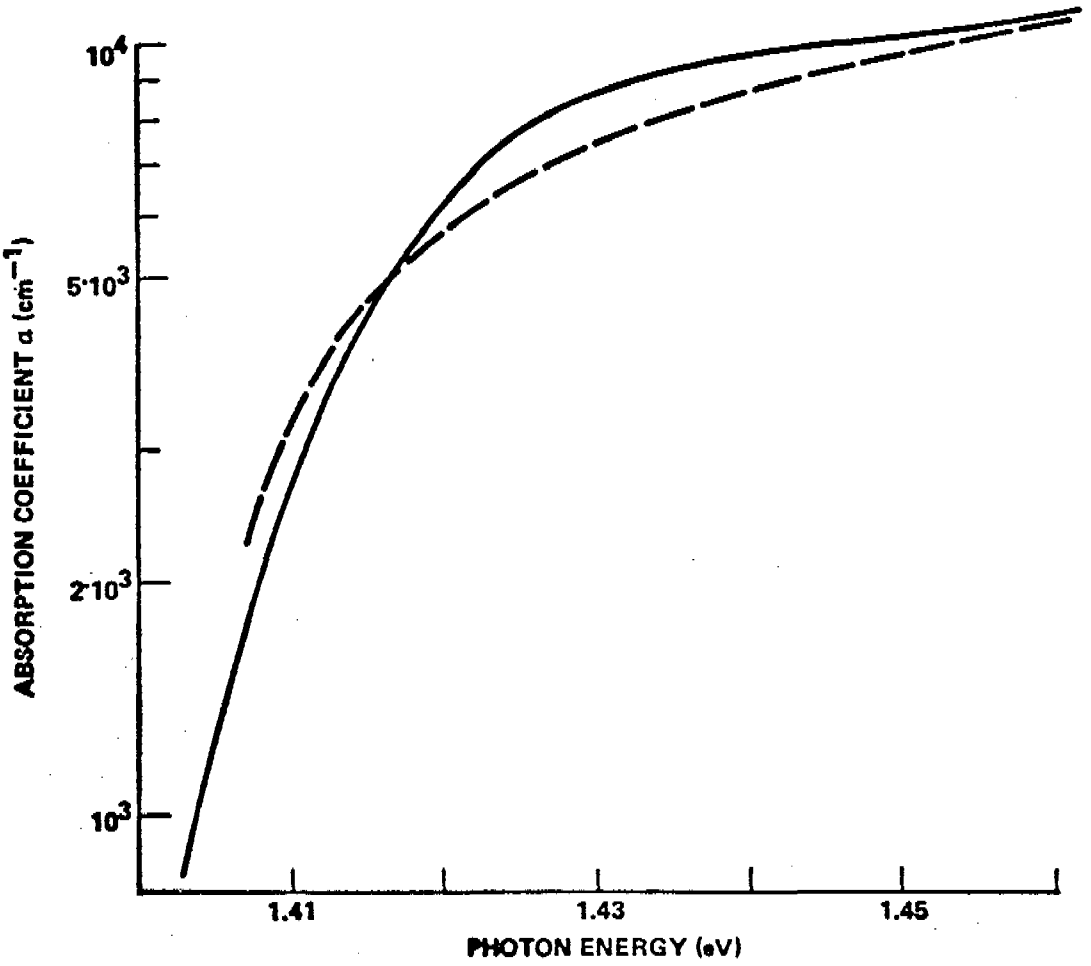


Fig. 7. Comparison between the experimental values of the absorption coefficient of GaAs at room temperature (ref. 15) (solid curve) and the Keldysh model calculation (dashed curve).

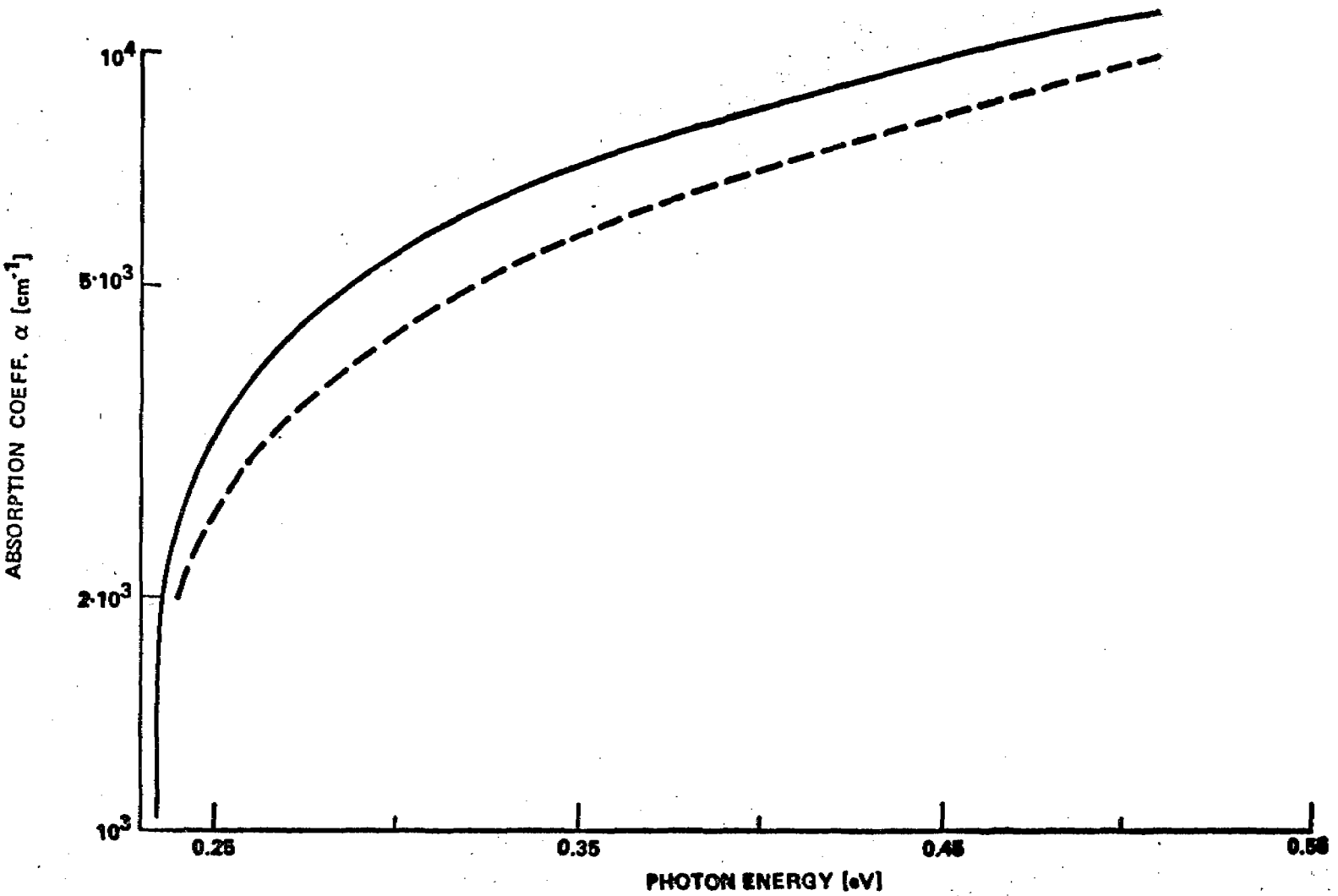


Fig. 8. Comparison between the experimental values of the absorption coefficient of pure InSb taken from ref. 12 (solid curve) and the Keldysh model calculation (dashed curve).

is rather good, particularly if we consider the large range of incident photon energies included in the comparison. A further discussion on the agreement of Keldysh model with the second-order perturbation theory is given in section 6.

5. SECOND-ORDER PERTURBATION MODELS. In this section, we consider two second-order perturbation models first proposed by Braunstein and Ockman[5] and by Basov et al.[6,7] As explained in section 3, in the second-order perturbation an intermediate state is required to complete the transition from the initial to the final state of the system perturbed by the radiation field. A proper accounting of intermediate states becomes important when their energy eigenstates are close to initial or final states of the system, or if they coincide with other real states encountered in practical materials such as exciton or impurity states. Therefore, the agreement between the calculated and the observed nonlinear absorption coefficients will depend largely upon the inclusion of the appropriate details of the energy band structure.

Braunstein and Ockman[5] consider vertical transitions between unperturbed parabolic bands. They assume that the only significant intermediate state is a higher conduction band designated with n in Fig. 9.

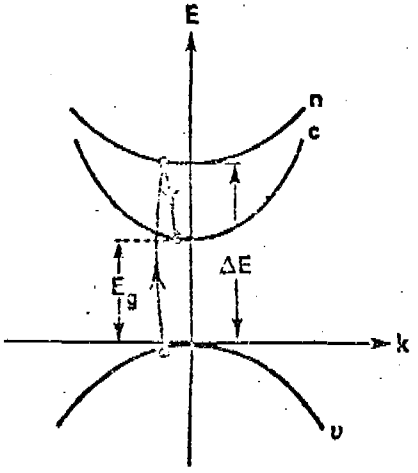


Fig. 9. The interconduction band transition model of Braunstein; v , c , and n refer to the valence, lowest conduction, and intermediate conduction bands, respectively. All energy bands are taken to be parabolic in the k-space.

Second-order perturbation theory yields the transition rate per unit volume for vertical transitions at a given value of k ; subsequently, the k-dependent probability is integrated over all values of k . For allowed-allowed transitions, the result is

$$w = \frac{|\mathcal{H}_{nc}|^2 |\mathcal{H}_{vn}|^2}{8\pi\hbar^4} \frac{m^{\frac{3}{2}}}{\sqrt{2}} \frac{(2\hbar\omega - E_g)^{\frac{1}{2}}}{(a_c + a_v)^{\frac{3}{2}}} \left(\Delta E - \hbar\omega + \frac{a_n + a_v}{a_c + a_v} (2\hbar\omega - E_g) \right)^{-2} . \tag{5.1}$$

In eq. (5.1), $\mathcal{H}_{nc}$ and $\mathcal{H}_{vn}$ are the matrix elements of the interaction Hamiltonian between the conduction c and the valence v bands, and the

intermediate conduction band n, respectively. The free electron mass is designated with m, and a_i $(i = c,v,n)$ is the reciprocal effective mass ratio m/m_i^* for the i-th band, $a_i = m/m_i^*$. The energy gap ΔE is the energy difference between the top of the heavy hole band (the only valence band included in this calculation) and the bottom of the intermediate conduction band.

The interaction Hamiltonian is taken to be of the form

$$\mathcal{H} = \frac{e}{mc} \vec{p} \cdot \vec{A} \quad .$$

We will evaluate the matrix elements of the interaction term by the so-called $\vec{k} \cdot \vec{p}$ method.[22] In this method, the matrix elements of the effective Hamiltonian are given by

$$\mathcal{H}_{j,\ell} = [E_j(\vec{k}_o) + \frac{\hbar^2}{2m} (\vec{k}^2 - \vec{k}_o^2)] \delta_{j,\ell} + \frac{\hbar}{m} (\vec{k} - \vec{k}_o) \cdot \vec{p}_{\ell,j} \quad . \tag{5.2}$$

Consider a solid consisting of two energy bands denoted by 0 and 1 each having an extremum in the reciprocal lattice space at $\vec{k}_o = 0$ and at corresponding momenta $\vec{p}_{01} = \vec{p}_{10}^* = \vec{p}$ taken to be isotropic. We choose the zero of energy axis to be at the top of the lower band 0 such that $E_o(\vec{k}_o) = 0$ and $E_1(\vec{k}_o) = E_g$ for the upper band, E_g being equal to the forbidden gap width. These assumptions restrict our treatment to solids with allowed direct transitions between the two bands. The Hamiltonian operator is then given by a 2×2 matrix shown below

$$\mathcal{H} = \begin{vmatrix} \frac{\hbar^2 k^2}{2m} & \frac{\hbar}{m} \vec{k} \cdot \vec{p} \\ \frac{\hbar}{m} \vec{k} \cdot \vec{p} & E_g + \frac{\hbar^2 k^2}{2m} \end{vmatrix} \quad . \tag{5.3}$$

Its diagonalization yields the following eigenvalues of the matrix (5.3)

$$E_{0,1}(k) = E_g/2 + \hbar^2 k^2/2m \pm \sqrt{E_g^2/4 + \hbar^2 k^2 p_{0,1}^2/m^2} \quad . \tag{5.4}$$

For small values of k, the square root in eq. (5.4) can be expanded in a power series about $E_g/2$. Retention of first-term only yields the following expression for the lower energy band

$$E_o(k) = \frac{\hbar^2 k^2}{2m} [1 - \frac{2p^2}{mE_g}] \quad , \tag{5.5}$$

and

$$E_1(k) = E_g + \frac{\hbar^2 k^2}{2m}\left[1 + \frac{2p^2}{mE_g}\right] \tag{5.6}$$

for the upper energy band. The effective reciprocal mass tensor of the carrier in n-th band is defined by

$$(m/m^*)_{n\alpha\beta} = \frac{m}{\hbar^2}\frac{\partial^2 E_n(k)}{\partial k_\alpha \partial k_\beta} \quad . \tag{5.7}$$

Because of the assumption of isotropic matrix elements in the Hamiltonian (5.3) of the two-band model, eqs. (5.5) to (5.7) yield scalar reciprocal effective masses for free carriers, $[m/m^*]_0 = 1 - 2p^2/(mE_g)$, and $[m/m^*]_1 = 1 + 2p^2/(mE_g)$ for the lower and the upper energy bands, respectively. We take the lower band 0 to be the highest valence band and the upper band 1 the lowest conduction band of a solid with a direct energy bandgap. The effective reciprocal masses of holes h in the valence v and electrons e in the conduction c bands are thus given by the scalar relations

$$[m/m_h^*]_v = 1 - 2p^2/(mE_g) \tag{5.8}$$

and

$$[m/m_e^*]_c = 1 + 2p^2/(mE_g) \quad , \tag{5.9}$$

respectively.

The band curvatures of the valence and the conduction bands are of opposite signs which accounts for the opposite signs of the p^2 term in the eqs. (5.8) and (5.9). Keeping in mind the opposite curvatures, we can rewrite eqs. (5.8) and (5.9) in absolute values to yield

$$|m/m_h^*| = 2p^2/(mE_g) - 1 \tag{5.8a}$$

and

$$|m/m_e^*| = 2p^2/(mE_g) + 1 \quad . \tag{5.9a}$$

From (5.8a) and (5.9a), one readily obtains $m/m^* = 4p^2/(mE_g)$. The energy-momentum relation needed to evaluate the matrix elements of the interaction Hamiltonian in terms of the one-electron effective mass approximation of the solid state theory is therefore given by

$$\frac{p^2}{2m} = \frac{E_g}{8}\left(\frac{m}{m^*}\right) \quad , \tag{5.10}$$

which is quite satisfactory for small values of k. We proceed in evaluating the matrix elements of eq. (5.1) with the help of eq. (5.10) and substitute the transition rate into eq. (2.2) for the case $\nu = 2$. The second-order nonlinear absorption coefficient is given by

$$\alpha_2 = \frac{\sqrt{2}\,\pi e^4}{c^2 \hbar\omega \epsilon_\infty} \frac{(\Delta E - E_g)\,\Delta E}{(\hbar\omega)^2} f_{nc}\, f_{vn}$$

$$\times \frac{(2\hbar\omega - E_g)^{\frac{1}{2}}}{\sqrt{m^*}} \left[\Delta E - \hbar\omega + \frac{a_n + a_v}{a_c + a_v}(2\hbar\omega - E_g)\right]^{-2} \quad . \tag{5.11}$$

The f_{nc} and f_{vn} denote the oscillator strengths for the transitions mediated by $\mathcal{H}_{vn}$ and $\mathcal{H}_{nc}$, respectively. In numerical calculations given in section 6, it is taken that $f_{nc} = f_{vn} = 1$. All parameters in eq. (5.11) are expressed in c.g.s. electrostatic units.

Equation (5.11) has been specialized for the case of single photon beam of frequency ω in the material (rather than two beams at two different frequencies ω_1 and ω_2). Braunstein and Ockman's expression for the nonlinear absorbance given by eq. (7) of ref. 5 must be multiplied by a factor of $1/(\epsilon_\infty^2)\left(m/(2m^*)\right)^2$ to obtain our eq. (5.11). The first term in the corrective factor is due to the different formulation of absorption coefficient in our eq. (2.2) as compared with eq. (2) of ref. 5, and the second term arises because we evaluated the matrix elements of allowed transitions in the $\vec{k} \cdot \vec{p}$ approximation rather than by employing eq. (4) of ref. 5. Our eq. (5.11) includes appropriate factors for spin orientation and photon multiplicity.

A generalization of Braunstein's calculation for anisotropic energy bands has been given by Hassan.[23] His results agree with our eq. (5.11) in the limit of zero anisotropy.

In Basov et al.[6,7] model, the two highest valence bands v_1 and v_2 and the lowest conduction band c are coupled by the radiation field. Furthermore, the intermediate states are conduction (or valence) intraband states as shown in Fig. 10.

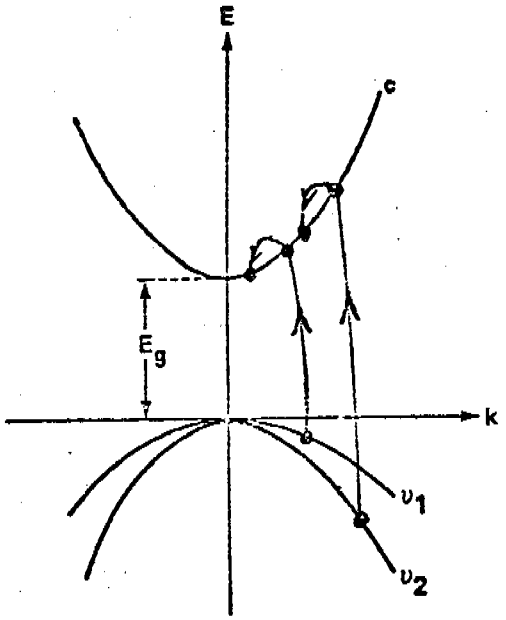


Fig. 10. The energy band model employed by Basov et al. Parabolic energy bands v_1, v_2 refer to the bound electrons in valence band, and c is the conduction band. Intermediate levels are provided by intraconduction and intravalence states. Only intraconduction band transitions are explicitly indicated.

Second order perturbation calculation of this model leads to the following expression for the second order absorption coefficient

$$\alpha_2^{(i)} = \frac{2 \cdot 2^9 \sqrt{2}\,\pi\, e^4 (m^*)^{\frac{1}{2}} (2\hbar\omega - E_g)^{\frac{5}{2}}}{\epsilon_\infty c^2 (\hbar\omega)^5} \frac{|(\hat{e} \cdot \vec{p})_{cv_i}|^2}{m^2} \qquad (5.12)$$

where $\hat{e}$ is the unit polarization vector of the incident radiation, $\vec{p}$ is the momentum operator of the electron and the index i = 1,2 refers to the initial state of the electron in the valence bands. Contrary to Basov et al., we use eq. (5.10) to evaluate the momentum matrix elements, and furthermore we average over all directions in the k-space. The total absorption coefficient must contain summation over both valence band contributions, $\alpha_2 = \Sigma_i \alpha_2^{(i)}$. Since each valence band contributes about equally, we need to calculate only $\alpha_2^{(1)}$. Therefore values listed in Table 7 of section 6 must be multiplied by a factor close to 2 to obtain the total second-order absorption coefficient in this model.

In our eq. (5.12), factor of 2 has been introduced to account for the spin degeneracy. This factor has been omitted in the original work by Basov et al. Further differences between our eq. (5.12) and Basov et al. expressions arise because of an incorrect numerical factor used in relating the flux intensity with the magnitude of vector potential in ref. 6. This gives a multiplicative correction of 1/16. Averaging the scalar product $|\hat{e} \cdot \vec{p}|^2$ over all directions in the k-space yields another multiplicative correction of 1/3. Hence, Basov et al. expression for the second-order absorbance in the context of eq. (6) of ref. 6 must be multiplied by a factor of 1/48.

This has been noted previously by Lee and Fan[24] and Fossum and Chang.[21] The need for this correction does not arise in connection with our eq. (5.12) because our absorption coefficient α_2 is intensity-independent and it was calculated from our eq. (2.2) rather than eqs. (6) and (8) of ref. 6, and the averaging over the k-space has been also performed in the calculations listed in section 6. A further difference between Basov et al. work and our eq. (5.12) arises from our use of the $\vec{k} \cdot \vec{p}$ method in evaluating the matrix elements of the interaction operator. Because of eq. (5.10), we evaluate

$$\frac{|(\hat{e} \cdot \vec{p})_{c-v_i}|^2}{m^2} = E_g/4m^* \quad ,$$

whereas Basov et al. use the relation $p^2/2m = E_g\, m^*/m$ in connection with eq. 5 and the relation $p^2/2m = 3E_g/4m_e^*$ in the experimental results section of ref. 6. In section 6, we report the numerical evaluation of eq. (5.12).

6. COMPARISON BETWEEN CALCULATED AND MEASURED ABSORPTION COEFFICIENTS. In this section, we compare absorption coefficients calculated from eq. (2.2) with available experimental data for the second-order nonlinear absorption. First, we evaluate the transition rates for photon multiplicities $\nu = 1$ to 3 in the perturbed valence and conduction band wavefunction model of Keldysh, eq. (4.10). For the second-order calculation, we also use the perturbation-theoretic approaches for the two different band models, eqs. (5.11) and (5.12).

In the Keldysh model calculation, it is necessary to evaluate the Dawson integral numerically by using the trapezoidal rule of integration. In general, the accuracy of the numerical integration is of the order of a few parts in 10^5. If it is desired to increase the accuracy of the integration, increase the number of intervals N in the program line 260. The computation time will increase accordingly.

The computer program listed in Table 4 has been written in a modified BASIC programming language for the HP 9830A calculator equipped with ROM's containing the mathematics package and the plotter control. This program (less lines 1000 and above, which contain the calculation of the absorption coefficient), with an appropriate modification of the plotting routine, has been also employed in plotting Figs. 1 to 6 of section 2.

Table 4. HP 9830A computer program used to calculate and to plot the wavelength dependence of absorption coefficients of the order 1 to 3 from eqs. (4.10) and (2.2). The computer system consists of extended memory core HP 9830A, built-in mathematics and plotter control ROM's, the HP 9866A printer and the HP 9862A plotter. Instructions for use are contained in lines 30 to 255.

```
30 REM THIS PROGRAM CALCULATES THE TRANSITION RATE PREDICTED BY KELDYSH
40 REM AND PLOTS VS PHOTON ENERGY IN ELECTRON VOLTS; KEY IN RUN AND FOLLOW
   INSTRUCTIONS
42 REM IT ALSO PLOTS ABSORPTION COEFF. VS PHOTON ENERGY IN ELECTRON VOLTS
50 REM NOTE; AT LEAST ONE INCREMENT MUST BE ASSIGNED
70 PRINT "INPUT PLOTTER SCALE, XMIN, XMAX, YMIN, YMAX"
71 PRINT
72 PRINT
80 INPUT S1, S2, S3, S4
90 SCALE S1, S2, S3, S4
95 XAXIS S3, S2/10, S1, S2
100 YAXIS S1, S4/2, S3, S4
115 FLOAT 3
120 PRINT "ENTER GAP IN EV, LAMBDA IN MICRONS"
130 PRINT "ENTER EL. FIELD IN UNITS OF MEGAVOLTS/M, MASS IN EL. MASS"
140 PRINT
150 INPUT G, L, E, M
160 PRINT
170 PRINT "IF YOU WANT TO INCREMENT GAP, KEY IN DELTAGAP; IF NOT KEY IN ZERO"
180 PRINT
190 INPUT G1
200 PRINT "IF YOU WANT TO INCREMENT LABMDA, KEY IN DELTALAMBDA; IF NOT, KEY
   IN ZERO
210 PRINT
220 INPUT L1
230 PRINT "IF YOU WANT TO INCREMENT EL. FIELD, KEY IN DELTAFIELD; IF NOT,
   ZERO"
240 PRINT
250 INPUT E1
252 PRINT "TO PLOT ABSORPTION COEFF. INPUT DIELECTRIC CONST (E2) IF NOT,
   ZERO"
253 PRINT
254 PRINT
255 INPUT E2
260 N=20
270 PRINT "GAP=";G,"LAMBDA=";L
280 PRINT "EL. FIELD=";E,"MASS=";M
290 PRINT "G1="G1;"L1="L1;"E1="E1
295 PRINT "S1="S1;"S2="S2;"S3="S3;"S4="S4
297 PRINT "E2="E2
300 PRINT
310 PRINT
320 A=8.76057E+36*(M/L)↑1.5/L
330 B=1.23752E-08*(E*L)↑2/(M*G)
340 X=0.806015*G*L*(1+B)
```

```
350 R1=INT(X+1)
360 Z=(2*R1-2*X)↑0.5
370 W1=A*(B/4)↑R1*EXP(2*R1*(1-B)-Z*Z)
380 Y=Z/N
390 P=(1+EXP(Z*Z))*Z/(2*N)
400 F=0
410 FOR I=1 TO N-1
420 F=F+EXP((I*Z/N)↑2)
430 NEXT I
440 W2=P+F*Z/N
450 W=W1*W2
460 IF G1=0 THEN 520
470 PLOT G,LGTW
490 IF G>5 THEN 630
500 G=G+G1
510 GOTO 320
520 IF L1=0 THEN 570
524 IF E2≠0 THEN 530
525 PLOT 1.2395/L,LGTW
527 IF E2=0 THEN 540
530 GOSUB 1000
540 IF L>21 THEN 630
550 L=L+L1
560 GOTO 320
570 IF E1=0 THEN 630
580 PRINT E,LGTW
600 IF E>50 THEN 630
610 E=E+E1
620 GOTO 320
630 END
1000 X=INT(G/(1.2395/L))+1
1010 A=2*X*1.2395/L*1.602E-19*754↑X*W/(SQRE2↑X*(E*1E+04)↑(2*X))
1020 PLOT 1.2395/L,LGTA
1030 RETURN
```

GaAs absorption coefficient for photon processes from 1 to 4 order calculated with the program listed in Table 4 is shown in Fig. 11. The material parameters used in this calculation are listed in Table 5. Experimental data for the absorption coefficient of GaAs are available for the first-order and second-order transitions only; they are entered as a dot, a vertical bar, and a circle in Fig. 11. It is seen that the agreement at 1.5 eV (one-photon linear absorption) and 0.94 eV (two-photon nonlinear absorption) between the Keldysh model prediction and the experiment is quite good. The sharp decline in absorption at the increase of the order of the photon multiplicity will be moderated by the contribution from the excitonic states. Since the calculation does not include exciton effects, one can expect that the experimental values will not agree with the predictions in the range of photon energies up to 0.1 eV less than that of the absorption edge for a particular order of photon multiplicity.

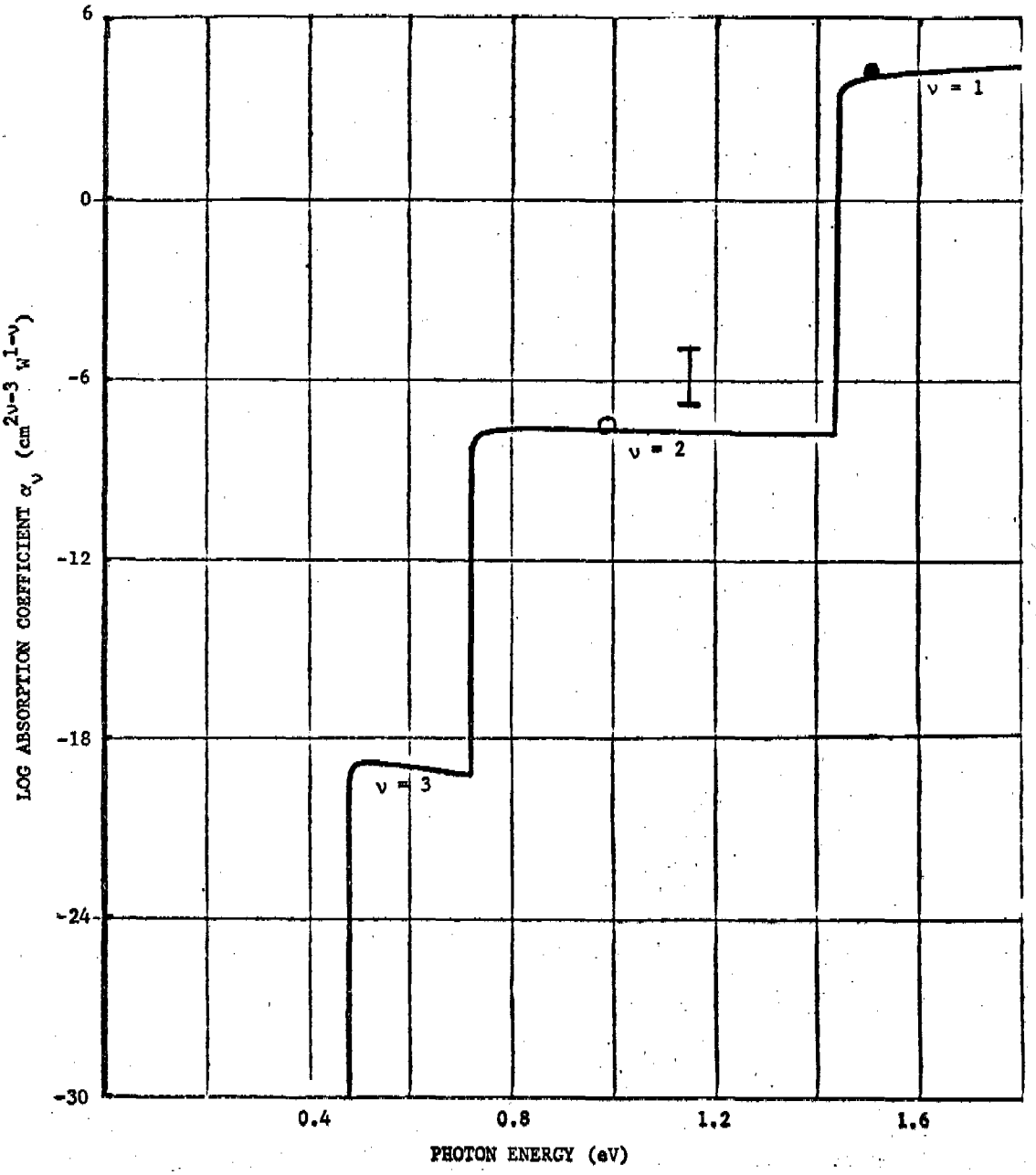


Fig. 11. Multiphoton absorption coefficient of GaAs at room temperature calculated from eqs. (4.10) and (2.2). See Table 5 for material parameters used in the calculation. The dimension of the absorption coefficient is given in units of $cm^{2\nu-3}\ W^{1-\nu}$, with ν designating the order of the absorption process. Experimental data are shown for 1.5 eV photon energy [16] (dot), 1.17 eV photon energy [6, 25 to 28] (vertical bar), and 0.94 eV [29 and 30] (circle).

Furthermore, impurities giving rise to energy levels within the forbidden gap would also increase the absorption coefficient. Undoubtedly, they also contribute to the large scatter of experimental values seen at 1.17 eV in Fig. 11. Because none of these contributions are accounted for in the Keldysh model, the predicted values of the absorption coefficients should be regarded as lower limits expected in the case of very pure materials. By inspection of Fig. 11, one can estimate that the second-order absorption will be significant at light fluxes above 10^6 W cm^{-2} and the third order - above 10^8 W cm^{-2} for the intrinsic nonlinear electronic transition processes in direct-bandgap materials. Absorption coefficients for several semiconductors of this type are listed in Tables 5 and 6 for fixed laser wavelengths of doubled Nd^{3+}-glass (0.53 μm), ruby (0.694 μm), Nd^{3+}-glass (1.06 μm), Nd^{3+}-YAG (1.318 μm), HF (2.8 μm), DF (3.9 μm), CO (5.3 μm), and CO_2 (10.6 μm).

Since the order of the transition is the dominant factor in determining the magnitude of the absorption, a large variation in the bandgap with the temperature implies a very large temperature dependence of the absorption coefficient. This is demonstrated in the last entry of Table 6 by listing absorption coefficients of PbS calculated at 300°K and 0°K. Because the perturbative methods have been carried out to the second order only, comparison of the second-order nonlinear absorption coefficient calculated for the three models of sections 4 and 5 is given in Table 7. For material parameters used in these calculations and comparison with the available experimental data, the reader is referred to Ref. 32. We note that the absorption coefficient calculated from the Keldysh model generally falls in-between the two perturbation models of Braunstein and Basov. However, the Basov model usually overestimates the absorption coefficient; therefore, the Keldysh model values yield currently the best estimate of the lower limit of the nonlinear absorption coefficients.

7. CONCLUSIONS. Comparison of theoretical models of Keldysh, Braunstein, and Basov in calculating the nonlinear absorption coefficients for direct band semiconductors with forbidden energy bandgaps between 2.6 to 0.15 eV show that the second order perturbation models of Braunstein and Basov differ in their prediction of the α_2 by almost three orders of magnitude. The Keldysh model prediction for α_2 lies between the two perturbation models and is usually slightly lower than the experimental value. A large scatter in reported experimental values of α_2 (perhaps attributable to the presence of impurities in real materials) hinders in selecting the best theoretical approach. However, generally it can be stated that if the interference terms in calculating transition rates are neglected, the absorption constants are underestimated (Braunstein and Ockman). Calculations allowing transitions from several bands (Basov) generally overestimate the absorption. The Keldysh method in which the effect of the optical electromagnetic field on the eigenfunctions of the unperturbed system is incorporated at the beginning of the calculation gives a slight underestimate of the two-photon absorption. It describes very well the band edge absorption of the one-photon process. Intuitively, one would concede that the perturbation of the eigenstates of the noninteracting system by the light intensities of interest is significant and that the perturbed eigenfunctions should be introduced at the onset of calculations as it is done in the Keldysh model. In addition, the Keldysh model is the only one currently available to

Table 5. Absorption coefficients of order (1) to (3) calculated from the Keldysh model for direct bandgap semiconductors of 2.6 to 1.2 eV bandgap. Material parameters listed in [18] and [31] were used; m* denotes the reduced effective pair mass and ϵ_∞ is the high frequency dielectric constant.

Material specifications				Absorption coefficient of order ν in units $cm^{2\nu-3}/W^{\nu-1}$ at a given wavelength λ in μm			
Material	Bandgap	m*	ϵ_∞	λ=0.530	λ=0.694	λ=1.06	λ=1.318
ZnSe	2.58 (300^oK)	0.132	5.9	$5 \cdot 10^{-9}$ (2)	$6.4 \cdot 10^{-9}$ (2)	$2.6 \cdot 10^{-21}$ (3)	$4.7 \cdot 10^{-21}$ (3)
CdS	2.53 (300^oK)	0.192	5.32	$4.7 \cdot 10^{-9}$ (2)	$6.1 \cdot 10^{-9}$ (2)	$1.8 \cdot 10^{-21}$ (3)	$3.4 \cdot 10^{-21}$ (3)
CdSe	1.74 (300^oK)	0.124	6.1	$6.5 \cdot 10^{4}$ (1)	$1.9 \cdot 10^{4}$ (1)	$1.7 \cdot 10^{-8}$ (2)	$1.65 \cdot 10^{-8}$ (2)
CdTe	1.50 (300^oK)	0.084	7.21		$3.4 \cdot 10^{4}$ (1)	$2.3 \cdot 10^{-8}$ (2)	$2.7 \cdot 10^{-8}$ (2)
GaAs	1.435	0.063	10.9		$2.7 \cdot 10^{4}$ (1)	$1.9 \cdot 10^{-8}$ (2)	$2.3 \cdot 10^{-8}$ (2)
InP	1.28 (300^oK)	0.062	9.56		$3.4 \cdot 10^{4}$ (1)	$2.6 \cdot 10^{-8}$ (2)	$3.17 \cdot 10^{-8}$ (2)

Table 6. Absorption coefficients of order (1) to (3) calculated from the Keldysh model for direct bandgap semiconductors of 0.8 to 0.15 eV bandgap. Material parameters listed in [18] and [31] were used; m* denotes the reduced effective pair mass and ϵ_∞ is the high frequency dielectric constant.

Material specifications				Absorption coefficient of order ν in units $cm^{2\nu-3}/W^{\nu-1}$ at a given wavelength λ in μm			
Material	Bandgap	m*	ϵ_∞	λ=2.8	λ=3.9	λ=5.3	λ=10.6
GaSb	0.8 (0^oK)	0.043	14.4	$8.7 \cdot 10^{-8}$ (2)	$3.4 \cdot 10^{-18}$ (3)		
GaSb	0.69 (300^oK)	0.043	14.4	$1.3 \cdot 10^{-7}$ (2)	$4.9 \cdot 10^{-18}$ (3)	$6.5 \cdot 10^{-18}$ (3)	
InAs	0.46 (0^oK)	0.019	11.8	$4.6 \cdot 10^{-7}$ (2)	$6.3 \cdot 10^{-7}$ (2)	$3.5 \cdot 10^{-7}$ (2)	
InSb	0.228	0.014	15.68			$1.4 \cdot 10^{3}$ (1)	$2.1 \cdot 10^{-6}$ (2)
PbTe	0.19 (0^oK)	0.011	3.69			$3.4 \cdot 10^{3}$ (2)	$1.6 \cdot 10^{-6}$ (2)
PbS	0.34 (300^oK)	0.36	18.5	$2.7 \cdot 10^{4}$ (1)	$1.5 \cdot 10^{-7}$ (2)	$2 \cdot 10^{-7}$ (2)	$1.1 \cdot 10^{-17}$ (3)
PbS	0.15 (0^oK)	0.034	18.5			$6.5 \cdot 10^{3}$ (1)	$3.2 \cdot 10^{-6}$ (2)

Table 7. Comparison of calculated two-photon absorption coefficient α_2 cm W^{-1} for direct bandgap semiconductors. Absorption coefficients were calculated from eq. (2.2); for transition rates, eqs. (4.10), (5.1), and (5.12) were used for Keldysh, Braunstein, and Basov models, respectively. Contribution from only one valence band is listed under Basov; the total absorption coefficient should include transitions from both valence bands and should be nearly twice as large.

Material	Wavelength in μm	α_2 [cm W^{-1}] calculated from model of: Keldysh	Braunstein	Basov
ZnSe	0.694	$6.4 \cdot 10^{-9}$	$2.2 \cdot 10^{-9}$	$4.5 \cdot 10^{-7}$
CdS	0.694	$6.1 \cdot 10^{-9}$	$2.2 \cdot 10^{-9}$	$4.4 \cdot 10^{-7}$
CdSe	1.06	$1.7 \cdot 10^{-8}$		$1.2 \cdot 10^{-6}$
	1.318	$1.65 \cdot 10^{-8}$		10^{-6}
CdTe	1.318	$2.7 \cdot 10^{-8}$	$9.6 \cdot 10^{-9}$	$1.8 \cdot 10^{-6}$
GaAs	1.06	$1.9 \cdot 10^{-8}$	$6.7 \cdot 10^{-9}$	$1.4 \cdot 10^{-6}$
	1.318	$2.3 \cdot 10^{-8}$	$8.25 \cdot 10^{-9}$	$1.5 \cdot 10^{-6}$
InP	1.06	$2.6 \cdot 10^{-8}$	$8.4 \cdot 10^{-9}$	$1.8 \cdot 10^{-6}$
InSb	10.6	$2.1 \cdot 10^{-6}$	$0.72 \cdot 10^{-6}$	$1.8 \cdot 10^{-5}$

calculate absorption coefficients higher than second order since there has been no published work carrying perturbation theory to a higher than second order in semiconductors. However, because intensities above 10^9 W cm^{-2} are required to attain a significant absorption in the third order in pure materials, such calculations may be superfluous for practical applications because of the onset of Drude absorption by free carriers created through the avalanche multiplication induced by the optical electric field. In fact, the notion that a single physical process can explain the effects of an intense laser pulse on a given optical material is too simplistic to hold because, for example, the multiphoton absorption and the avalanche ionization are two competing processes having a different time dependence.

On the other hand, this contribution demonstrates that it is important to include properly the band parameters of the solid state into the calculation. In particular, if contributions from excitonic states and impurity levels should be included, a summation over all corresponding elements of the interaction Hamiltonian must be incorporated into the calculation of absorption coefficients. The outline of such a procedure is given in Ref. 33. Furthermore, up to the third order, it is sufficient to use the $\vec{p} \cdot \vec{A}$ interaction term in the electric dipole approximation.

8. ACKNOWLEDGMENTS. The authors are thankful to Dr. C. C. C. Lee for pointing out inconsistencies in the original Basov et al. work and for his comments on the contributions of excitonic states to the nonlinear absorption. Also, thanks are due to Mr. J.E. Williams who has programmed several calculations reported in this work.

9. REFERENCES.

1. V. I. Bredikhin, M. D. Galanin, and V. N. Genkin, Usp. Fiz. Nauk 110, 3 (1973) [Sov. Phys.-Usp. 16, 299 (1973)].

2. J. M. Worlock, Two-photon Spectroscopy, in Laser Handbook, Vol. 2, Arecchi and Schulz-DuBois, eds., North-Holland (1972).

3. S. Bakos, Multiphoton Ionization of Atoms, Adv. in Electronics and Electron Physics, Vol. 36, L. Marton, ed., Academic Press (1974).

4. L. V. Keldysh, Zh. Eksp. Teor. Fiz. 47, 1945 (1964) [Sov. Phys.-JETP 20, 1307 (1965)].

5. R. Braunstein and N. Ockman, Phys. Rev. 134, A499 (1964).

6. N. G. Basov, A. Z. Grasyuk, I. G. Zubarev, V. A. Katulin, and O. N. Krokhin, Zh. Eksp. Teor. Fiz. 50, 551 (1966) [Sov. Phys.-JETP 23, 366 (1966)].

7. N. G. Basov, A. Z. Grasyuk, V. F. Efimkov, I. G. Zubarev, V. A. Katulin, and J. M. Popov, Proceedings of the International Conference on the Physics of Semiconductors, Kyoto 1966, J. Phys. Soc. Japan 21 Suppl., 277 (1966).

8. M. Goeppert-Mayer, Naturwissenschaften 17, 932 (1929); Annalen der Physik 9, 273 (1931).

9. W. Franz, Zeitschr. f. Naturforschg. 13a, 484 (1958).
L. V. Keldysh, Zh. Eksp. Teor. Fiz. 34, 1138 (1958) [Sov. Phys.-JETP 7, 768 (1958)].

10. H. B. Bebb and A. Gold, Phys. Rev. 143, 1 (1966).

11. M. H. Weiler, M. Reine, B. Lax, Phys. Rev. 171, 949 (1968).

12. I. M. Catalano, A. Cingolani, and A. Minafra, Solid State Comm. 16 (1975).

13. J. H. Yee, Phys. Rev. B3, 355 (1971).

14. C. H. Lee, "Interaction of Intense Picosecond Light Pulse with Materials," Progress Reports, Defense Documentation Center #AD741391 and AD 750414 (1972).

15. T. S. Moss and T. D. F. Hawkins, Infrared Phys. 1, 111 (1961).

16. M. D. Sturge, Phys. Rev. 127, 768 (1962).

17. I. Kudman and T. Seidel, J. Appl. Phys. 33, 771 (1962).

18. D. Long, Energy Bands in Semiconductors, J. Wiley, N.Y. (1968).

19. E. Kartheuser, in Polarons in Ionic Crystals and Polar Semiconductors, ed., J. T. Devreese, North-Holland Publ. (1972).

20. G. W. Gobeli and H. Y. Fan, Phys. Rev. 119, 613 (1960).

21. H. J. Fossum and D. B. Chang, Phys. Rev. B8, 2842 (1963).

22. J. Callaway, Quantum Theory of the Solid State (Academic Press, New York, 1974), Part A, pp. 248-252.

23. A. R. Hassan, Il Nuovo Cimento 70B, 21 (1970).

24. C. C. Lee and H. Y. Fan, Appl. Phys. Letters 20, 18 (1972).

25. Y. A. Oksman, A. A. Semenov, V. N. Smirnov, and O. M. Smirnov, Fiz. Tekh. Poluprov. 6, 731 (1972) [Sov. Phys. Semicond. 6, 629 (1972)].

26. C. C. Lee and H. Y. Fan, Phys. Rev. B9, 3502 (1974).

27. A. Z. Grasyuk, I. G. Zubarev, V. V. Lobko, Yu. A. Matveets, A. B. Mirnov, and O. B. Shatberashvili, ZhETF Pis. Red. 17, 584 (1973) [Sov. Phys. JEPT Lett. 17, 416 (1973)].

28. S. Jayaraman and C. H. Lee, Appl. Phys. Letters 20, 392 (1972).

29. D. A. Kleinman, R. C. Miller, and W. A. Nordland, Appl. Phys. Letters 23, 243 (1973).

30. J. M. Ralston and R. K. Chang, Appl. Phys. Letters 15, 164 (1969).

31. J. I. Pankove, Optical Processes in Semiconductors (Prentice Hall, Englewood Cliffs, N.J., 1971).

32. S. S. Mitra, L. M. Narducci, R. A. Shatas, Y. F. Tsay, and A. Vaidyanathan, Appl. Optics, September 1975, in press.

33. C. C. C. Lee, Two Photon Absorption and Second Harmonic Generation in Semiconductors, doctoral dissertation, Purdue University, May 1974 (unpublished).

A TIME-DEPENDENT QUANTIZED NATURAL COLLISION COORDINATES METHOD*

Norman M. Witriol
Physical Sciences Directorate
US Army Missile Research, Development and Engineering Laboratory
US Army Missile Command
Redstone Arsenal, AL 35809

ABSTRACT. In the reactive or nonreactive molecular collisional energy transfer problem, the quantum mechanical natural collision coordinate method has been shown to be useful in predicting the distribution of the energy of the reactants into the translational and vibrational energies of the products. This method has only been employed in a time-independent formalism. However, time-dependent studies of this problem via the solution of the time-dependent Schrödinger equation in the original coordinates have revealed some interesting insights into the dynamical effects taking place. With the goal of obtaining the advantages of using natural collision coordinates in a time-dependent method, a study of time-dependent quantum mechanical canonical point transformations has been initiated. One finds that by proceeding to the Klein-Gordon equation, a similar formalism to that used in handling time-independent quantum mechanical canonical point transformations can be employed. Applying this method, one obtains the transformed Hermitian coordinate momenta and the transformed Hermitian Hamiltonian. The Klein-Gordon equation and the transformed wavefunction, in the transformed space, are explicitly displayed.

---

*The full text of this paper will be published elsewhere.

COMPUTER SIMULATION OF THE INTERMEDIATE BALLISTIC ENVIRONMENT OF A SMALL ARM

Csaba K. Zoltani
Fluid Mechanics Branch
Applied Mathematics & Sciences Laboratory
USA Ballistic Research Laboratories
Aberdeen Proving Ground, MD 21005

ABSTRACT. In this paper the results of computation of the gas flow in the presence of the moving projectile in the intermediate ballistic range are presented. The mathematical model is based on the assumptions of cylindrical symmetry with unsteady, inviscid, compressible motion of a one component gas. The flow variables at the muzzle plane were varied continuously in accordance with the prediction of an interior ballistic calculation.

Velocity, pressure and temperature time histories, as well as pressure contours of the flow field were computed up to real times of 35μsec. The fine meshing employed in the difference technique enabled the resolution of significant flow details, including the backward facing shock at the base of the projectile. The calculated results are in satisfactory agreement with available analytical and experimental data.

1. INTRODUCTION. About 100μsec. before the base of the projectile clears the plane of the muzzle of an M-16 rifle, the first indication of gas motion out of the gun tube becomes apparent. This is in the form of a weak air shock, traveling at approximately M=1.5 with the pressure rise of the order of 16 atmospheres. About 10μsec. later, the gases which leaked around the projectile, while the latter is still in the barrel, appear starting to form the characteristic gas cloud. First evidence of a vortex formation around the lip of the muzzle is also observed at this time. Once the base of the projectile clears the muzzle, the combustion products stream out of the gun tube, and envelope the projectile, which at a velocity of 990 $[\frac{m}{s}]$, is moving slower than the gases. Near the centerline of the muzzle the pressure is around 600 atmospheres and temperature approximately 2500°K. A backward facing shock forms on the base of the projectile, which as the projectile accelerates, gradually weakens and disappears as soon as the powder bell is exited by the bullet.

The shock due to the powder bell behaves in the far field like a point explosion with the center of energy deposition located about half a caliber in front of the gun tube. The motion of this shock front is described quite well by self-similar calculations. At about 100μsec. after shot ejection, as the projectile

leaves the intermediate ballistic range, the powder bell is approximately 17.5 calibers in radius and 22 calibers in length, measured downstream from the muzzle. A typical shock Mach number at this time is around M ~ 2.0.

The presence of muzzle devices complicates the flow field considerably due to shock refraction and wave interaction. The muzzle brake used in this study, though an idealized version of the standard M-16 muzzle attachment, exhibits all the salient characteristics of the flow field.

The basic features of muzzle flow have been understood since the pioneering researchers of Cranz [1] in the 1920's. Numerical calculation of the flow field have been attempted only recently, notably by Oswatitsch [2], Taylor [3] and the group at Dahlgren [4]. The latter two are not quite satisfactory, as explained elsewhere [5], which motivated the current study.

II. Mathematical Analysis of the Flow Field.

A. The Flow Model. The calculations are based on a two-dimensional time dependent flow model. The geometry of the flow field (and of the muzzle devices) is assumed to be circular symmetric with respect to the axis of the gun barrel. The flowing medium is assumed to be a one-component, inviscid, non-reacting gas.

At the beginning of the calculations the projectile is placed 1.5mm (i.e., three mesh widths) in front of the muzzle and assumed to have the "muzzle velocity" of 990m/s.

The initial conditions for the gas are ambient pressure (1 atm) and temperature (300K), and zero velocity everywhere, except between exit plane of the muzzle and the projectile. At that location we assume high enthalpy and sonic (choked) flow conditions, constant across the muzzle opening. (The initial gas velocity is, e.g., 1210 m/s). The variation of these conditions at the exit plane with time are taken from one-dimensional interior ballistics calculations [6]. Because the exit flow is sonic, such calculations are independent of the intermediate ballistics events and constitute the time dependent boundary condition at the muzzle for the present calculations.

The projectile's acceleration during the intermediate ballistics phase is computed from pressure forces in accordance with Newton's law of motion. Viscous drag forces are neglected because the working medium is assumed to be inviscid for the present model.

A schematic of the computing mesh and initial conditions is shown in Figure 1.

The governing equations for the gas are in cylindrical coordinates are as follows:

$$\frac{\partial \rho}{\partial t} + \frac{\partial(r\rho u)}{\partial r} + \frac{\partial(\rho v)}{\partial z} = 0 \qquad (1)$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial r} + v\frac{\partial u}{\partial z} + \frac{1}{\rho}\frac{\partial p}{\partial r} = 0 \qquad (2)$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial r} + v\frac{\partial v}{\partial z} + \frac{1}{\rho}\frac{\partial p}{\partial z} = 0 \qquad (3)$$

$$\frac{\partial E}{\partial t} + u\frac{\partial E}{\partial r} + v\frac{\partial E}{\partial z} + \frac{1}{\rho}\left\{\frac{1}{r}\frac{(pru)}{\partial r} + \frac{(pv)}{\partial z}\right\} = 0 \qquad (4)$$

E is the specific total energy of the fluid, defined by $E = 1/2(u^2 + v^2) + I$ where I is the specific internal energy.

Equations (1-4), the basis of our analysis, represent the conservation of mass, momentum in the two co-ordinate directions, and energy respectively. The Noble-Abel equation of state, valid for most propellant reactions products, makes the above system determinate. It is

$$p = \frac{RT}{\frac{1}{\rho} - \beta}\,\frac{1}{M_i} \qquad (5)$$

where β is the co-volume, for the calculation at hand having a value of 0.001 $[\frac{m^3}{kg}]$, and M_i is the mole mass of the gas with R the universal gas constant. The temperature is calculated from:

$$T = \frac{(\gamma-1)\cdot I}{R} M$$

with M the molecular weight of the gas and I is the specific internal energy.

B. The Boundary Conditions. The boundaries were handled in the following manner: along the axis of symmetry, reflection was used, i.e. a row of imaginary cells was defined in which all the quantities have the same values as in the last row of cells in the flow field except for the normal component of the velocity, the sign of which reversed in the imaginary cell. The same technique was used along all the solid boundaries, that is, along the muzzle walls, the projectile and the muzzle device.

Outflow along the free boundaries had to be guaranteed. To prevent wave reflection, which was observed as the shock intersected the last computational cell, a new method of handling the free boundaries was introduced, see Figure 2. In essence, it involves subtracting out mass in the last three cells, ensuring that

a negative pressure gradient exists there facilitating the outflow.

The conditions at the muzzle were specified by an interior ballistics calculation [6]. For the first 50μs after the projectile's ejection they were as follows:

$$u = 1210 - 1.340 \times t$$
$$v = 0.000$$
$$p = (486.0 - 0.561 \times t) \times 10^5$$
$$\rho = 40.95 - 0.092 \times t$$

where t is specified in microseconds.

C. The Finite Difference Technique. The calculation proceeds in two steps. The first step is pure Lagrangian in that the convective terms are dropped from the equations of motions. The pressure gradients are approximated by leapfrog differencing with the pressure of the cell boundary being an average of the two neighboring cells in each co-ordinate direction. The result of the first step is a set of intermediate quantities, no physical significance being abscribed to their values. It should be noted that in the energy equation, the velocities used are averages of the initial and the intermediate values so that in fact I_j, the intermediate energy of the flow at location j, uses a different velocity base than the momentum equation. Although this makes the scheme in a sense inconsistent, the original developers of the algorithm [7] claimed that this procedure was necessary to ensure energy conservation.

Phase two of the algorithm updates the flow quantities to the final time by taking account of the transport terms neglected in phase one. The continuity equation uses donor cell differencing, i.e. the mass flow across the side of a cell is calculated by taking the density of the donor cell and the velocity is a weighted average of the velocity of the material flowing from the donor and the recipient cell. This velocity is obtained by a Taylor series expansion about the cell boundary. Analogous procedures are followed for the momentum and the energy equations. The time step was determined by the usual CFL, region of influence criterion with a factor 0.25, i.e.

$$\Delta t < 0.25 \frac{\min(\Delta r, \Delta z)}{a + \sqrt{u^2 + v^2}}$$

where a is the local sound speed. The calculations were carried out on a non-uniform mesh with 150 cells in the axial direction and 60 cells in the radial direction. Computational experience showed that cells could not have an aspect ratio, L/D greater than 1.5 and maintain stability, and in addition, cells had to be closely

spaced near the plane of muzzle, otherwise wild oscillations were observed. In the axial direction then, the first three cells, containing high pressure gases were spaced at 5.0×10^{-4} [m] It was at this distance that the base of the projectile was put at time t=0. Dowstream from that location, the cell size was gradually increased to 1.0×10^{-3} [m], then maintained constant further downstream.

Typical time steps, which of course are a function of mesh spacing, were around: 10^{-7} sec. Running times for 35μsec real time were of the order of 4 hours on BRLSC II.

III. Results and Discussion. The code employed in these calculations is a revised version of DORF[8], modified to include moving boundaries [9]. Considerable improvement of [9] was necessary, however, before acceptable results were obtained.

Figure 3 and 4 show pressure contours of the flow field at early times. The isobars, proceeding from the outermost inward, are set at 0.1215, 0.2, 0.4, 0.8, 1.0, 5.0, 10.0, 20.0, 50.0 [MPa] respectively. A shock is indicated where these are closely spaced, all around the periphery of the gas cloud. Due to the inherent diffusion of the finite difference technique, the shock is smeared, being 0.003 [m] in thickness. At t = 15.41μsec. (Figure 4) the interaction with the muzzle device is clearly evident as well as the gas motion around the tip of the projectile. The pressure at the shock front is typically 6 atmospheres. The next graph, Figure 5, shows the pressure along the stagnation streamline (i.e. axis of the barrel) for cycles 100, 200, 300 corresponding to real times of 5.50, 10.58, 14.41 [μsec]. At cycle 100 the pressure drops from 4×10^{7} [Pa] at the nozzle to around 9.0×10^{6} [Pa] toward the base of the projectile. A slight rise at the base of the projectile indicates the presence of a backward facing shock. This shock comes about because the gases are still moving faster than the projectile, the base of which appears like a blunt body to the flow. The projectile pushes against the ambient at its tip producing the lump in the gas pressure there corresponding to a bow shock. At succeeding cycles the results are qualitatively analogous, with the pressure continuously dropping at the base of the projectile due to the lateral expansion of the gas. Figure 6 shows the radial pressure profile in a plane at 0.0015[m] in front of the muzzle at t=15.41μsec proceeding in the radial direction. The profile shows the expected steep pressure drop in the direction away from the muzzle, followed by a rise to around 5.0×10^{5} [Pa] at the blast wave.

In Figure 7 we have plotted the radial temperature profile at t = 34.29μsec. in the plane 1.5 mm in front of the muzzle. The interesting point here is the temperature rise,

at a distance of 34 x 10^{-3}[m] from the axis of symmetry, indicating the presence of the shock.

The next figure, Figure 8, shows the axial velocity along the axis of symmetry at three different times. The gas velocity increases downstream from the muzzle as the gases expand, reaching a value of slightly in excess of 2500[m/s] before steeply dropping at the base of the projectile. Near the nose of the projectile, one can observe a steep drop to zero, the gases having been only slightly affected by the motion of the projectile at these early times.

A number of checks were made on the results described. These consisted of Rankie-Hugoniot conditions in the normal and radial directions to see if the conservation laws were obtained. Deviations of up to 25% were observed, but in view of the fact that the RH conditions pertain to steady flow, while here we one dealing with an unsteady situation, the agreement was judged to be satisfactory.

The shock stand-off distance on the base of the projectile was also calculated. Using the correlation of reference [10],

$$\frac{\delta}{R} = \frac{1.03}{\sqrt{\frac{\rho}{\rho_\infty} - 1}} \qquad (6)$$

where δ is the stand-off distance, R the characteristic dimension and ρ_∞ the density of the gas ahead of the shock, agreement within 10% of that predicted by formula (6) was found.

Finally, the calculated pressure was checked against experimental values. Measurements are considered to be unreliable due to the hostile environment of the flow field and the fact that the insertion of a probe into the flow field will alter its structure. Figure 9 shows the correlation between theory and experiment.

In conclusion then, we have presented the results of the simulation of the flow field, for early times at the muzzle of a small caliber weapon. The numerical results are in satisfactory agreement with our qualitative understanding of the flow.

REFERENCES

1. Cranz, C., Lehrbuch der Ballistik, Springer Verlag, Berlin, 1926.

2. Oswatitsch, K., Intermediate Ballistics, DVL Report No. 358, Aachen, 1964.

3. Taylor, T.D., Calculation of Muzzle Blast Flow Field, Technical Report 4155, Picatinny Arsenal, Dover, NJ (1970).

4. Moore, G.R., Maillie, F.H, Soo Hoo, G., Calculation of 5"/54 Muzzle Blast and Post Ejection Environment on Projectile, NWL Technical Report TR-3000, January 1974.

5. Zoltani, C.K. Numerical Simulation of the Muzzle Flow Field with a Moving Projectile. Proceedings, First International Symposium on Ballistics, American Defense Preparedness Association, Washington, D.C. 1974, pp.II-127-151.

6. Celmins, A.K.R., Theoretical Basis of the Recoilles Rifle Interior Ballistics Code RECRIF, BRL Report in preparation.

7. Rich, M., A Method for Eulerian Fluid Dynamics. LAMS-2826 (1962).

8. Payton, D.M., The DORF User's Manual. AFWL-TR-70-60, July 1970.

9. Farr, J.L., Traci, R.M. A User's Manual for SAMS. BRL Contractor Report 162, June 1974.

10. Serbin, H., Supersonic Flow Around Blunt Bodies. J. Aeronautical Sciences 25, 58-59 (1958).

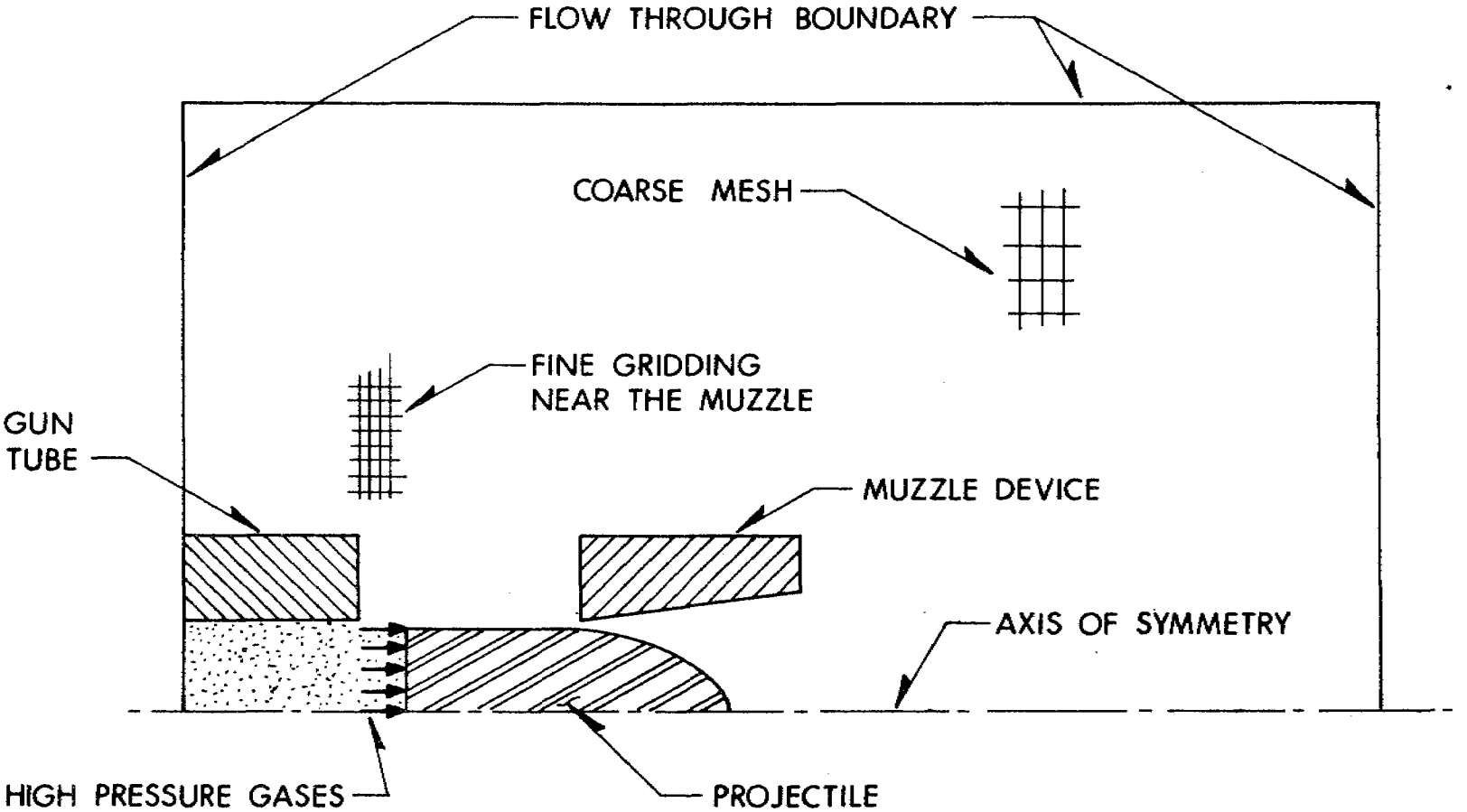


Figure 1. Schematic of the Computational Region at t = 0.000

OUTFLOW BOUNDARY CONDITIONS

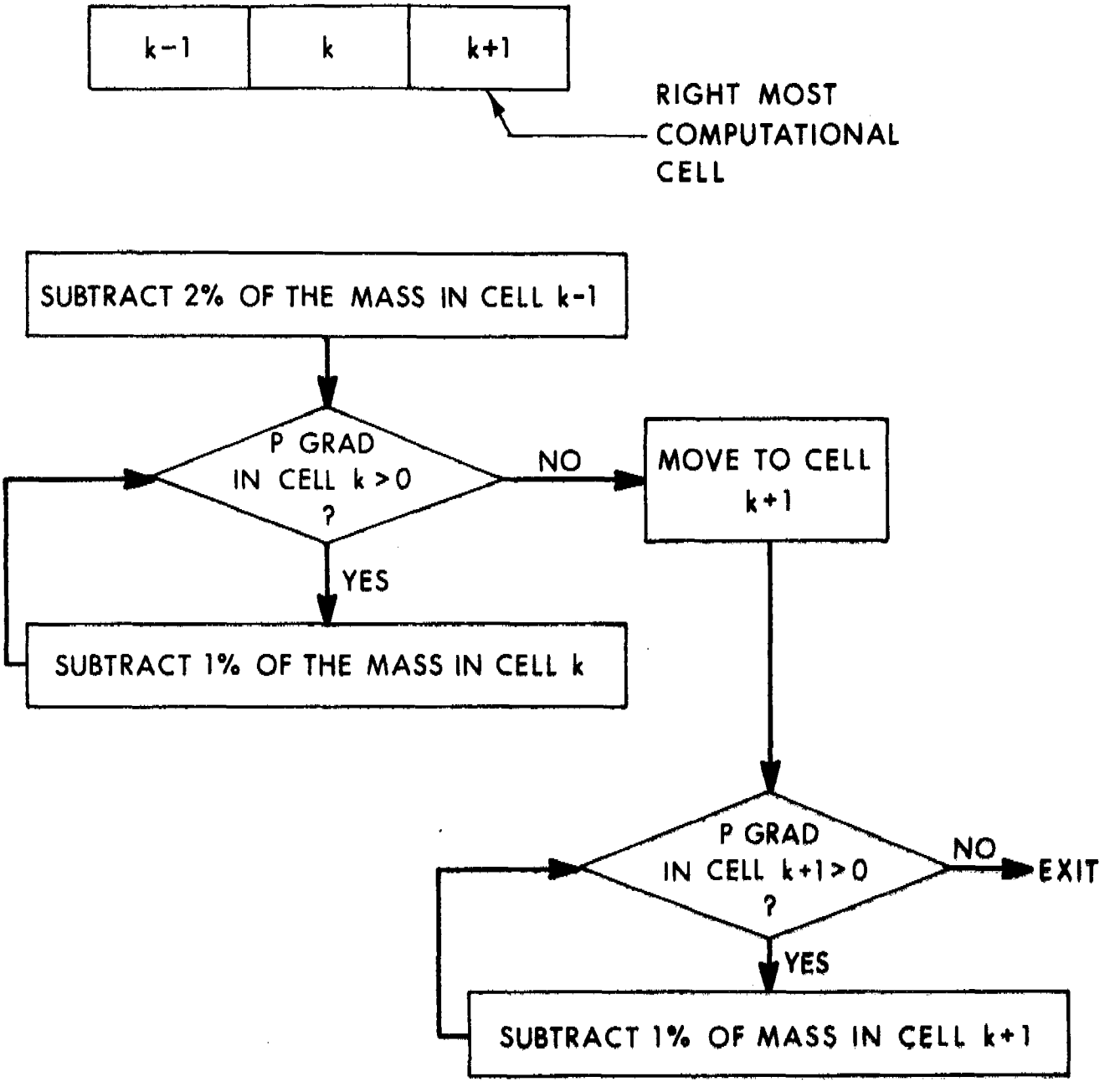


FIGURE 2. THE SCHEME FOR THE OUTFLOW BOUNDARIES.

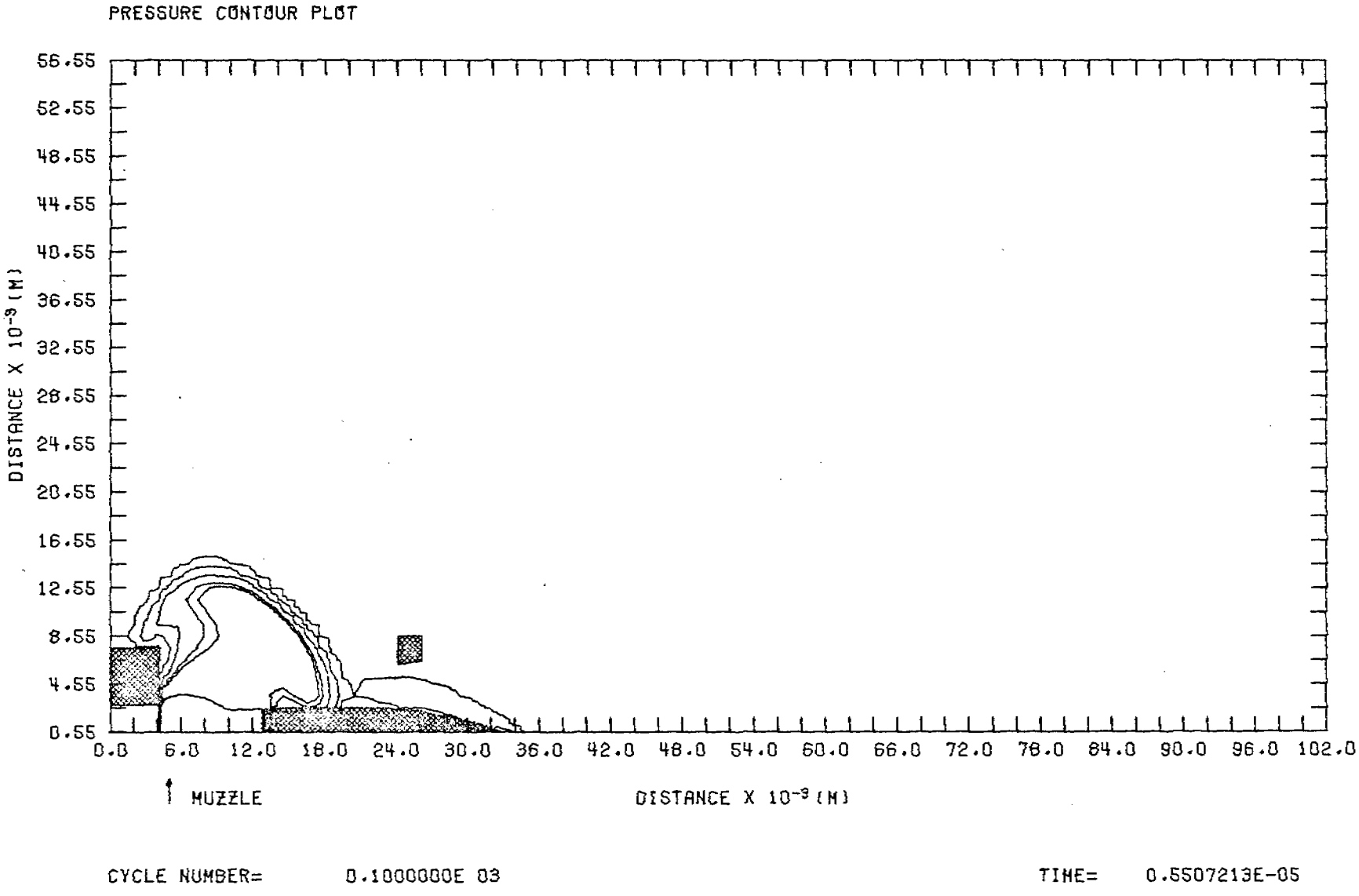


FIGURE 3. FLOW FIELD WITH MOVING PROJECTILE SHOWING LINES OF CONSTANT PRESSURE.

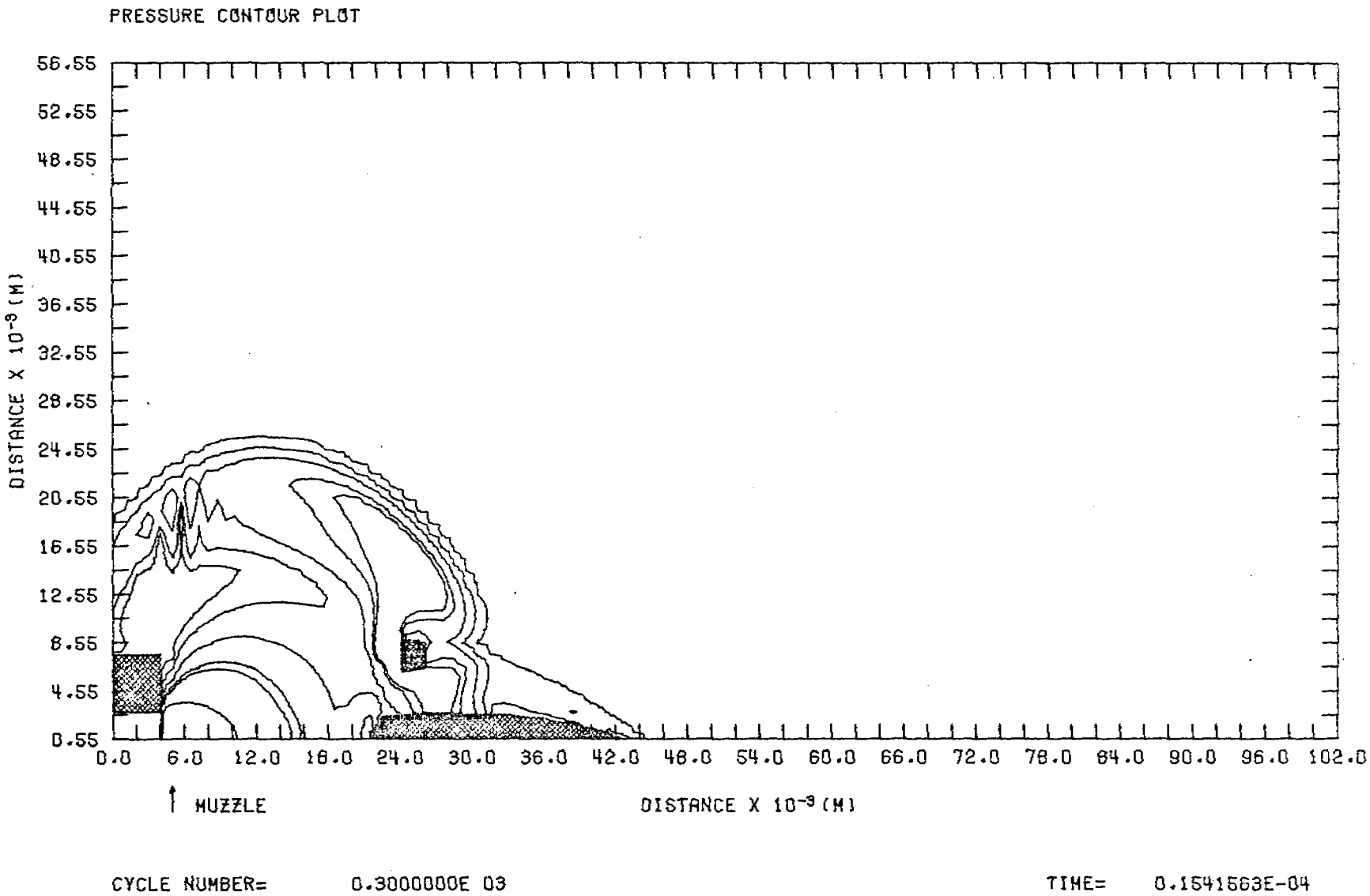


FIGURE 4. THE FLOW FIELD AT T = 15.41 μSEC.

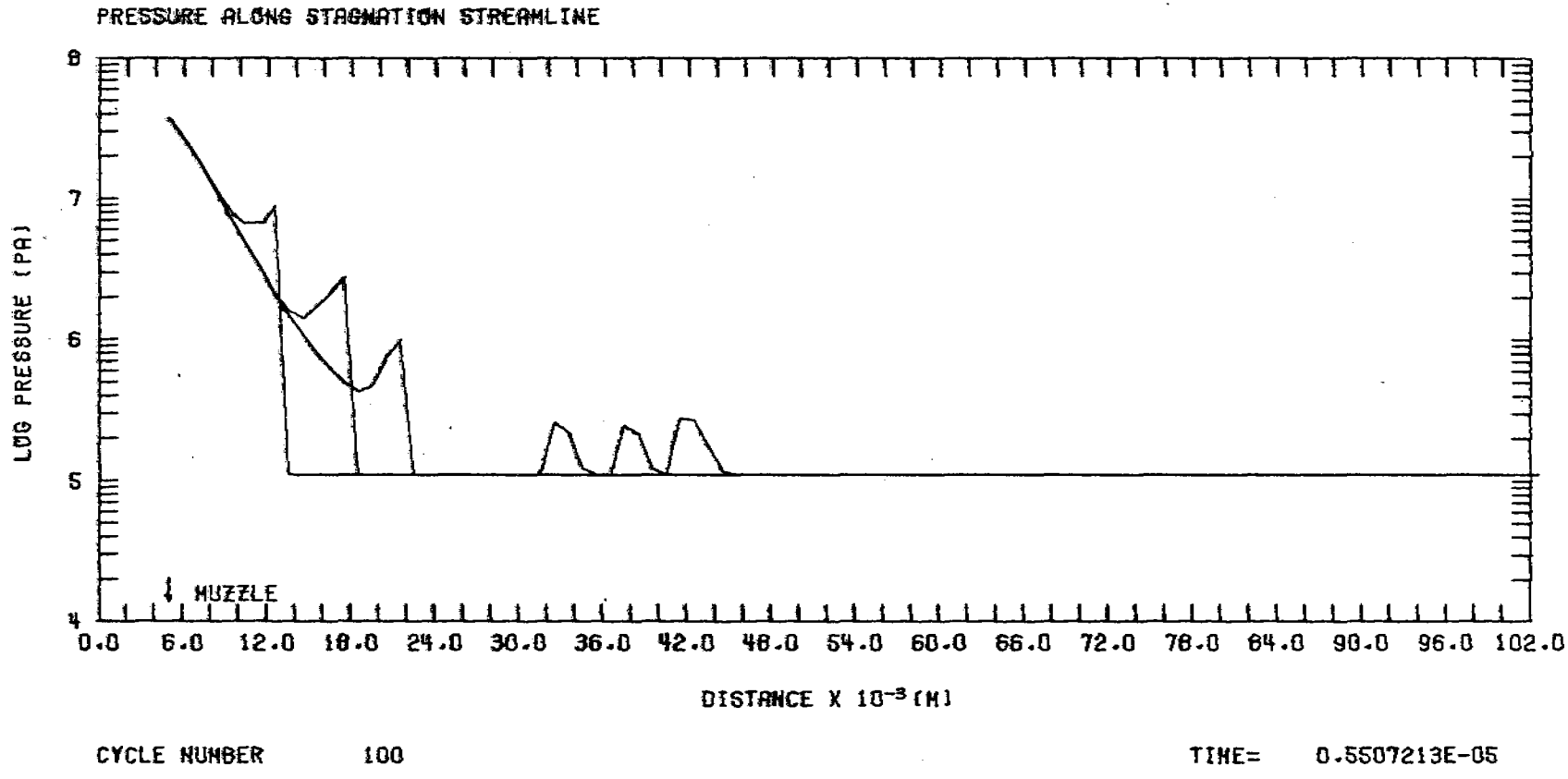


FIGURE 5. PRESSURE ALONG THE STAGNATION STREAMLINE. THE CYCLE NUMBER ARE 100, 200 AND 300 CORRESPONDING TO T = 5.50, 10.58 AND 15.41 μS.

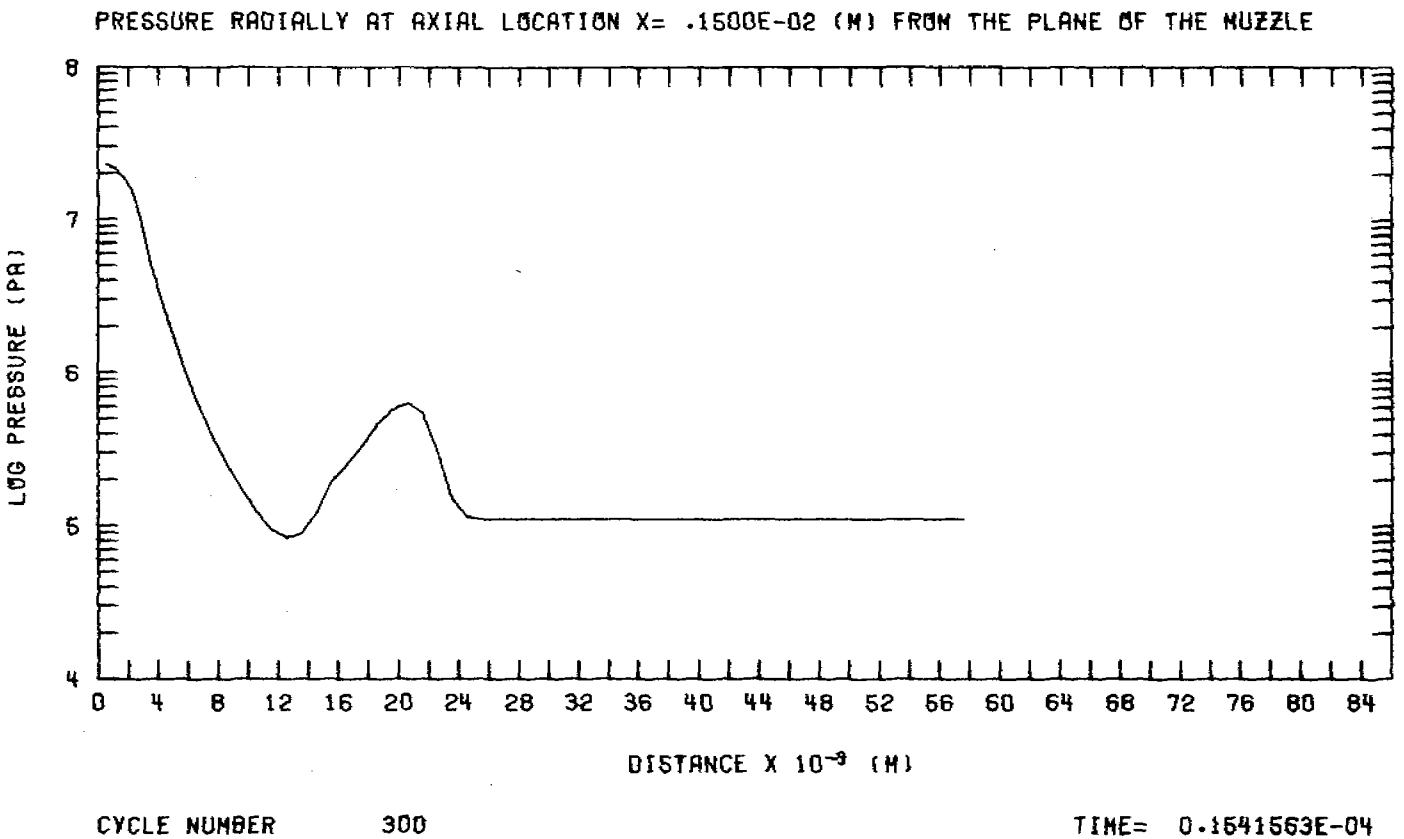


FIGURE 6. PRESSURE RADIALLY IN A PLANE AT 0.0015(M) FROM THE MUZZLE.

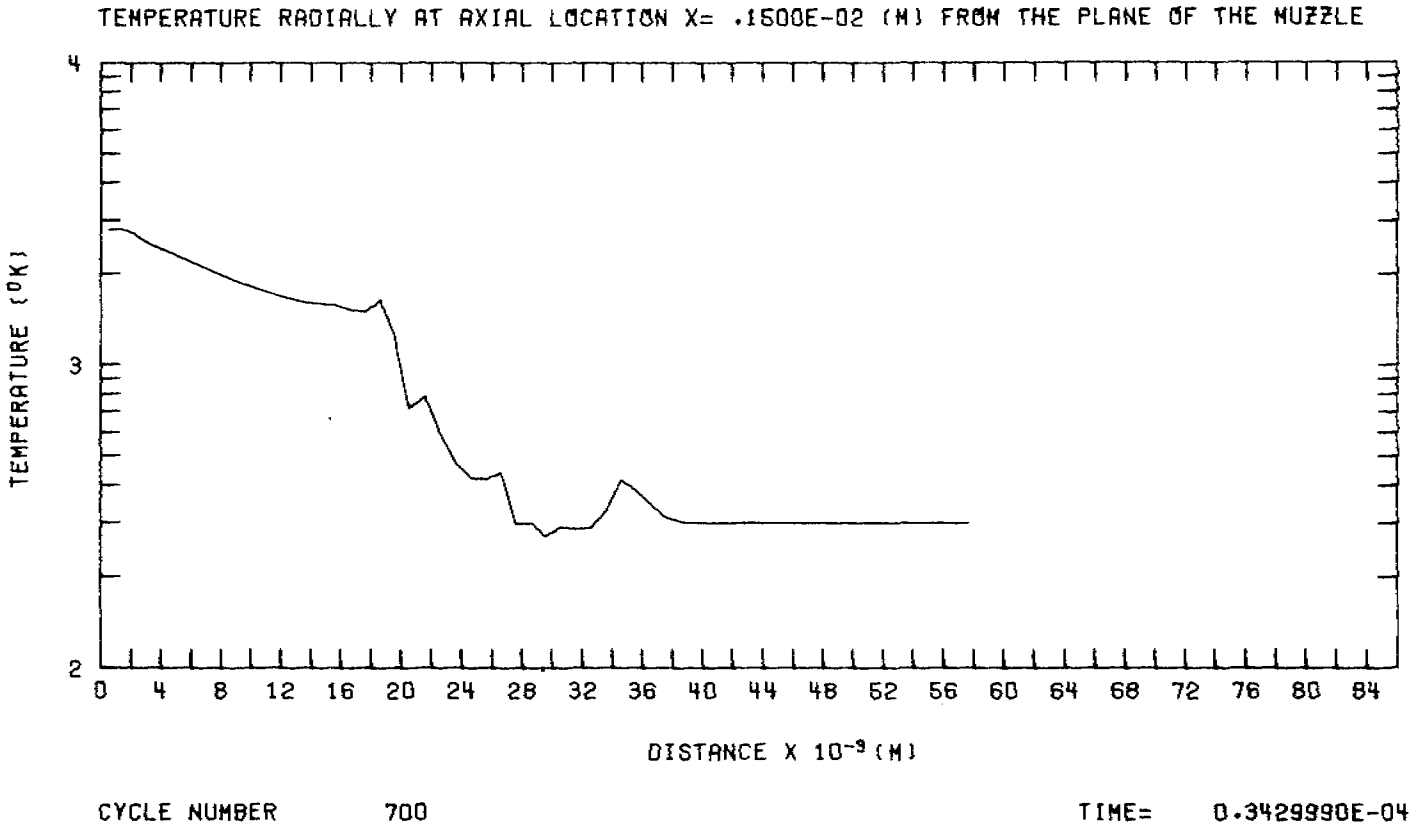


FIGURE 7. TEMPERATURE RADIALLY IN A PLANE AT 0.0015(M) IN FRONT OF THE MUZZLE.

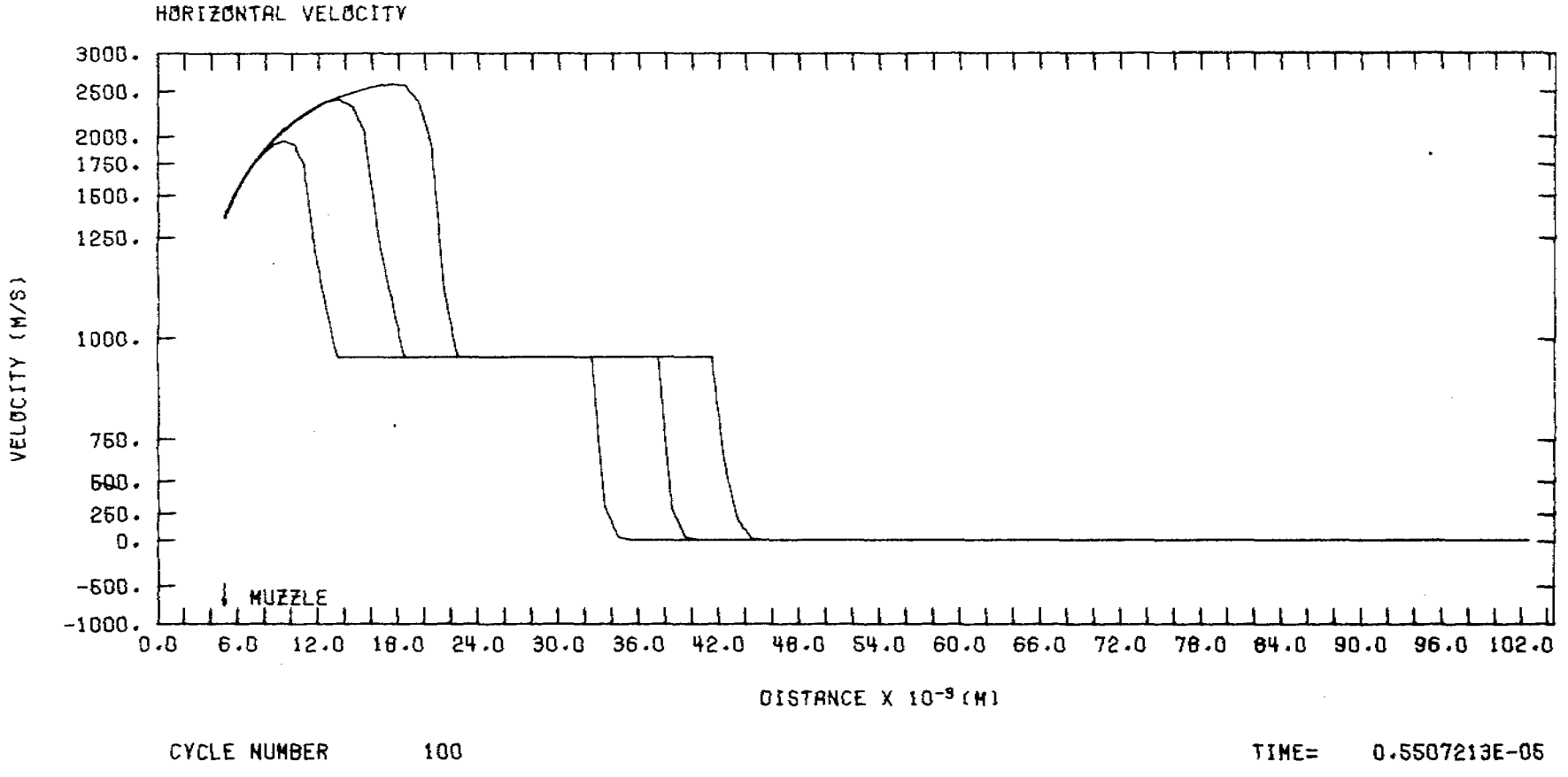


FIGURE 8. THE HORIZONTAL VELOCITY ALONG THE AXIS OF SYMMETRY AT CYCLES 100, 200, 300, CORRESPONDING TO T = 5.50; 10.50; AND 15.41 μS.

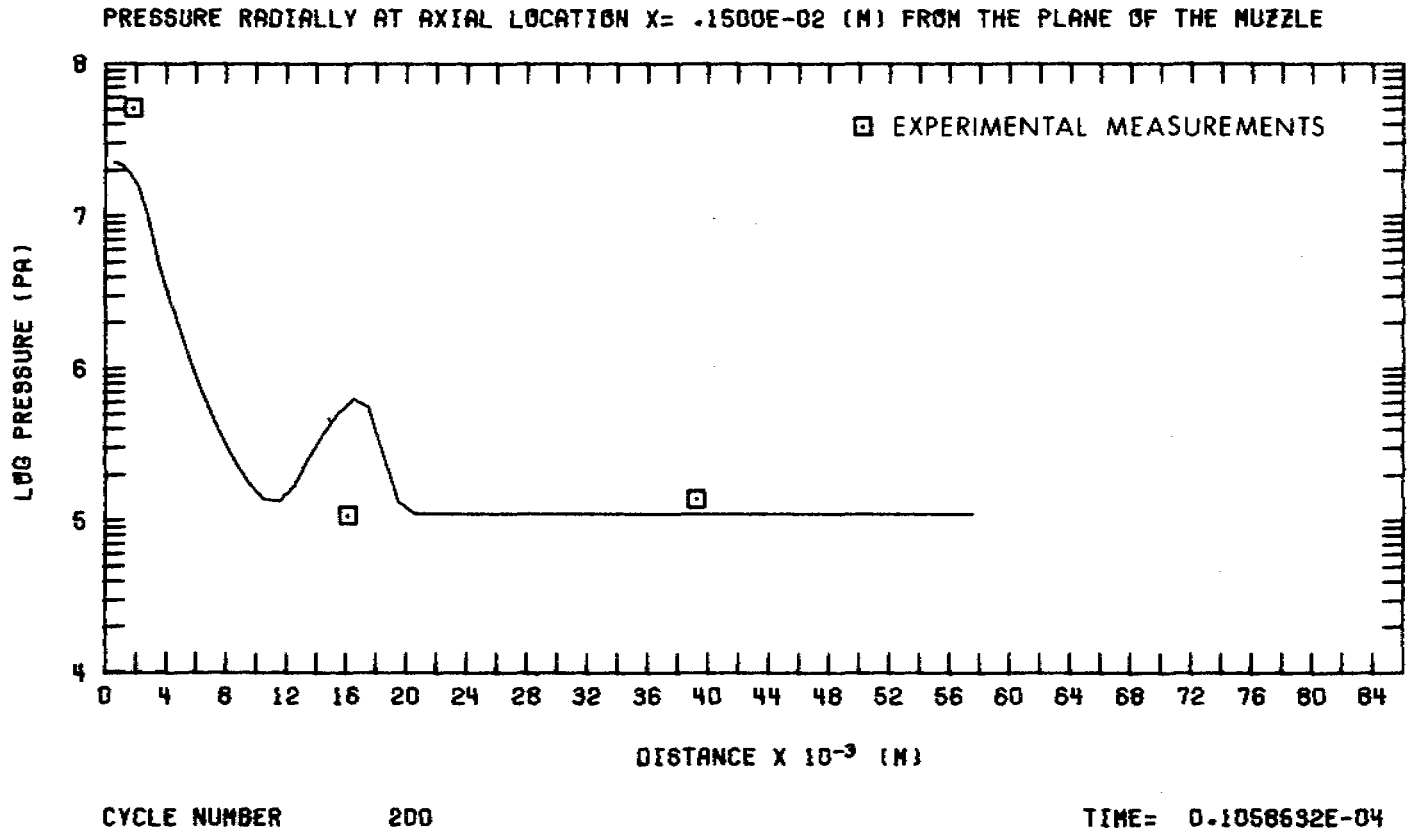


FIGURE 9. COMPARISON BETWEEN COMPUTED AND MEASURED PRESSURE VALUES.

GENERALIZED SHOCK WAVE PHYSICS:

ELECTROMAGNETIC AND SECOND SOUND SHOCKS

Paul Harris
Concepts and Effectiveness Division
Nuclear Development and Engineering Directorate
Picatinny Arsenal
Dover, New Jersey 07801

ABSTRACT The steady shock velocities for finite amplitude electromagentic and second sound disturbances are derived. A generalized form for the shock velocity of an arbitrary disturbance, in terms of "jump" conditions in stimulus and response parameters, is suggested. That generalized form is compared with the derived shock velocities, and with the shock velocity for a finite amplitude mass density disturbance.

I. INTRODUCTION

A steady state finite amplitude compressional mass density disturbance is characterized by a shock velocity, U , given by

$$U = \sqrt{\frac{\rho_i \{P\}}{\rho_f \{\rho\}}} \tag{1}$$

where P dentoes pressure, ρ is mass density, the subscripts i and f denote initial (unshocked) and final states respectively, and the curly brackets indicate that the quantity is to be evaluated across the shock front (surface).

$$\{P\} \equiv P_f - P_i$$

The form of Eq. (1) has been known for more than a century[1].

In sections II and III we will derive expressions for the shock velocity associated with finite amplitude electromagnetic and second sound disturbances. Those expressions will be seen to be of a form similar to Eq. (1), and lead us to suggest that a generalized finite amplitude disturbance propagates at a shock velocity given by

$$U = \sqrt{\phi \frac{\{A_1\}\{A_2\}\{A_3\}\cdots}{\{B_1\}\{B_2\}\{B_3\}\cdots}} \quad , \tag{2}$$

where A_i denotes the variable assoicated with the applied i^{th} stimulus, B_i the variable characterizing the corresponding response, and ϕ is a quantity assoicated with dimensionality.

II. ELECTROMAGNETIC SHOCK WAVES

Let us consider the two time-varying Maxwell's equations for electromagnetic field propagation in a medium.

$$\vec{\nabla}\times\vec{H} = \frac{4\pi}{c_o}\vec{J} + \frac{1}{c_o}\frac{\partial\vec{D}}{\partial t} \quad , \tag{3}$$

$$\vec{\nabla}\times\vec{E} = -\frac{1}{c_o}\frac{\partial\vec{B}}{\partial t} \quad , \tag{4}$$

where gaussian units have been used, c_o is the velocity of light in vacuum, and the symbols have their usual meaning[2].

We restrict ourselves to spatial variation only in the x_3 (or z) direction. This corresponds to the so called "one-dimensional" strain configuration of mass density shock wave problems. Thus

$$-\frac{\partial E_2}{\partial x_3} = -\frac{\partial B_1}{\partial t} \quad \text{and} \quad \frac{\partial E_1}{\partial x_3} = -\frac{1}{c_o}\frac{\partial B_2}{\partial t} \quad , \tag{5}$$

$$-\frac{\partial H_2}{\partial x_3} = \frac{4\pi}{c_o}J_1 + \frac{1}{c_o}\frac{\partial D_1}{\partial t} \quad \text{and} \quad \frac{\partial H_1}{\partial x_3} = \frac{4\pi}{c_o}J_2 + \frac{1}{c_o}\frac{\partial D_2}{\partial t} \quad . \tag{6}$$

The steady shock can be described by employing $\frac{\partial}{\partial t} = U\frac{\partial}{\partial x_3}$ in Eqs. (5) and (6). The second of Eqs. (5) and the first of Eqs. (6) then give

$$dE_1 = -\frac{U}{c_o}dB_2 \quad , \tag{7}$$

$$-dH_2 = \frac{4\pi}{c_o}J_1 dx_3 + \frac{U}{c_o}dD_1 \quad , \tag{8}$$

where dependence on only a single independent variable allows total differentials to be used.

Integrating Eqs. (7) and (8) across the shock front, which may be of finite thickness, and solving for U, gives

$$U = \sqrt{c_o^2\,\frac{\{E_1\}[\{H_2\} + 4\pi c_o^{-1}\int_{sh}J_1 dx_3]}{\{D_1\}\{B_2\}}} \tag{9}$$

For the special case $J_1 = 0$, Eq. (9) is the form of Eq. (2); E_1 and H_2 are identified as stimuli, and D_1 and B_2 the corresponding response variables. The integral in Eq. (9) is to be taken across the shock front. Physically, J_1 enters as it does because a surrounding magnetic field is always associated with a current. Eq. (9) is well known. Indeed, it is the basis for important effects in pulse shaping and transmission line technology[3].

The existence of a shock wave is generally associated with nonlinear material properties. Although we have not explicitely exhibited such nonlinearities in arriving at Eq. (9), they are in principal present. For example, D_1 can have quadratic dependence on E_1, and/or field dependent breakdown can contribute to J_1.

In actual practice the existence of nonlinear material properties is a necessary but not a sufficient condition for the existence of a shock wave. In order for an electromagnetic shock to exist, it is necessary that the nonlinearity be such that U increases as the magnitude of E_1 and/or H_2 increases. Suffice it to say that systems do exist which have nonlinearities of the type necessary for the formation of electromagnetic shock waves[3].

Regardless of whether or not the shock condition exists, Eq. (9) gives the velocity of that finite amplitude disturbance being propagated. As it must, Eq. (9) yields $U = c_o$ for propagation in a vacuum.

III. SECOND SOUND SHOCK WAVES

Second sound is typically a low temperature phenomenon. It is defined as the coherent propagation of a thermal disturbance. In pure second sound there is no associated mass density (or pressure) disturbance. The more everyday (say room temperature) mode of thermal propagation is via incoherent diffusive processes; second sound conditions are achieved when the frequency content of the thermal disturbance is of the order of the reciprocal of the thermal relaxation time for the system in question[4,5].

Second sound was first observed[6] in liquid helium (helium II), and has since been observed in non-superfluid materials[7]. Finite amplitude second sound can give rise to second sound shock which has been observed[8,9] in liquid helium. In this paper we will derive the velocity of a steady second sound shock in liquid helium.

Within the context of the two fluid model for superfluidity the equations[9] for the conservation of mass, the conservation of momentum, and the conservation of entropy, become

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial z}(\rho_s v_s + \rho_n v_n) = 0 , \tag{10}$$

(11a) $$\rho_n \frac{\partial v_n}{\partial t} = -\frac{\rho_n}{\rho}\frac{\partial P}{\partial z} - \rho_s s \frac{\partial T}{\partial z} ,$$

(11b) $$\rho_s \frac{\partial v_s}{\partial t} = -\frac{\rho_s}{\rho}\frac{\partial P}{\partial z} + \rho_s s \frac{\partial T}{\partial z} ,$$

(12) $$\frac{\partial}{\partial t}(\rho s) + \frac{\partial}{\partial z}(\rho s v_n) = 0 .$$

In Eqs. (1) - (12) the physical parameter s denotes specific entropy, T is temperature, v is the particle (flow) velicity, and the subscripts s and n refer to the superfluid and normal components respectively. ρ_s and ρ_n are constrained by $\rho = \rho_s + \rho_n$, and only the normal component is a carrier of entropy. Second sound gives the additional constraint

(13) $$dP = d\rho = 0 .$$

As in the electromagnetic case we employ $\frac{\partial}{\partial t} = U \frac{\partial}{\partial z}$. Thus from Eq. (11a)

(14) $$U dv_n = -\frac{1}{\rho} dP - \left(\frac{\rho_s}{\rho_n}\right) s dT .$$

Integrating Eq. (14) with $dP = 0$ gives

(15) $$v_n = -\frac{1}{U}\int_{shock} \left(\frac{\rho_s}{\rho_n}\right) s dT .$$

Substituting the steady state shock condition into Eq. (12) gives

(16a) $$U d(\rho s) + d(\rho s v_n) = 0 , \text{ or}$$

(16b) $$U ds + v_n ds + s dv_n = 0 .$$

Combining Eqs. (14), (15), and (16b) yields

(17) $$\left[U - \frac{1}{U}\int_{sh} \left(\frac{\rho_s}{\rho_n}\right) s dT \right] ds = \frac{s^2}{U}\left(\frac{\rho_s}{\rho_n}\right) dT ,$$

with the final result

(18) $$U = \sqrt{\frac{\int_{sh}\left(\frac{\rho_s}{\rho_n}\right) s^2 dT}{\int_{sh}\left[1 - \frac{1}{U^2}\int_{T_o}^{T}\left(\frac{\rho_s}{\rho_n}\right) s dT\right] ds}} .$$

While use of Eq. (18) is limited in that Eqs. (11) represent only the linearized version of superfluidity[10], it is important to realize that the same procedure utilized here is applicable to the nonlinear case; the result would be a more complicated version of Eq. (18).

To begin comparing Eqs. (2) and (18) we neglect the inner integral in the denominator if Eq. (18), and we average the integrand in the numerator. The result should apply to finite though small amplitude shocks and is given by

$$U \simeq \sqrt{\left[\frac{\rho_s}{\rho_n} s^2\right]_{avg} \frac{\{T\}}{\{s\}}} \quad . \tag{19}$$

Upon identification $\phi = (\rho_s \rho_n^{-1} s^2)_{avg}$ we have the desired form of Eq. (2). T is recognized as the stimulus, s the response, and second sound is to be viewed as the propagation of a coherent entropy disturbance.

In the limit of an infinitesimal disturbance Eq. (19) reduces to the accepted[9] expression for second sound velocity, U_{lim},

$$U_{lim}^2 = \frac{\rho_s}{\rho_n} s^2 \left(\frac{\partial T}{\partial s}\right)_P = \frac{\rho_s}{\rho_n} \frac{s^2 T}{c_P} \quad , \tag{20}$$

where c_P is the specific heat at constant pressure.

The next degree of analysis of Eq. (18) is to consider the integral in the numerator. This represents a situation similar to the integral over the current density in Eq. (9), and is present because ϕ is a strong function of the stimulus T. Physics and dimensional analysis would presumably have allowed one to use Eq. (2) in order to guess at the form of the numerator in Eq. (18). The form of the inner integral in the denominator of Eq. (18) is somewhat less obvious.

Again not all nonlineatities are sufficient to cause a positive (Δ T) disturbance to shock up. The existence of such a shock depends upon the detailed behavior of $(\rho_s \rho_n^{-1} s^2)$ as a function of T.

By using the pure second sound conditions of Eq. (13), we have neglected the higher order physics associated with thermal expansion and the Grüneisen parameter. The usual literature[11] derivation of the second sound shock velocity is perturbative (second order in Δ T) and includes the mentioned higher order effects. The form of the second sound shock velocity thus arrived at is

$$U = U_{lim} + \beta(v_s + v_n) \quad . \tag{21}$$

The higher order effects can be included in the approach leading to Eq. (18) by retaining the terms in $d\rho$ and dP. It is interesting to note that Eq. (19) can be placed in a form similar to Eq. (21), although without the corresponding physical content, by using

$$\{T\} = \frac{\partial T}{\partial s}\{s\} + \frac{1}{2}\frac{\partial^2 T}{\partial s^2}\{s^2\} \ . \tag{22}$$

IV. DISCUSSION

We have shown that an intelligent use of Eq. (2) is a good starting point for guessing at the shock velocity of an arbitrary disturbance in a complicated system. If we could not have guessed at the denominator of Eq. (18), we could certainly have written an empirical form for $\{B\}$ of Eq. (2) which might have suggested the correct solution.

Among the more esoteric nonlinear systems to which one might wish to apply Eq. (2) are those of quantum field theory[12] (propagation of a probability amplitude), and magnetohydrodynamic (MHD) shocks in gases[13] and solids[14]. A propagating distrubance in a nonlinear medium can, in general, give rise to shock formation, and its study may be of interest for a wide variety of physical, biological or social systems.

REFERENCES

1. See, for example, the discussion in sections 51-55 of R. Courant and K.O. Friedrichs, Supersonic Flow and Shock Waves (Inerscience, New York, 1951).

2. J.D. Jackson, Classical Electrodynamics (John Wiley and Sons, New York, 1962).

3. I.G. Katayev, Electromagnetic Shock Waves (Iliffe Books Ltd., London, 1966).

4. M. Chester, Phys. Rev. 131, 2013 (1963).

5. M. Weymann, Amer. J. Phys. 35, 488 (1967).

6. V. Peshkov, J. Phys, U.S.S.R. 8, 131 (1944).

7. Second sound has been observed, for example, in bismuth. V. Narayanamurti and R.C. Dunes, Phys. Rev. Letters 28, 1461 (1972).

8. D.V. Osborne, Proc. Phys. Soc, A, 64 114 (1951).

9. K.R. Atkins, Liquid Helium (Cambridge Univ. Press, Cambridge, 1959).

10. For a survey of some of the nonlinear theories, see J.G. Daunt and R.S. Smith, Rev. Mod. Phys. 26, 172 (1954).

11. I.M. Khalatnikov, An Introduction to the Theory of Superfluidity (W.A. Benjamin, Incl, New York, 1965). Chapter 13.

12. W. Heisenberg, Physics Today 20, 27 (May 1967).

13. Plasma Physics in Theory and Application, W.B. Kunkel, editor (McGraw-Hill, New York, 1966).

14. A.C. Baynham and A.D. Boardman, Plasma Effects in Semiconductors (Taylor and Francis Ltd., London, 1971).

ON RIEMANN'S INVARIANT AND SHOCK IMPEDANCE OF SOLIDS

Y. K. Huang
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. Using the Riemann invariant and Rankine-Hugoniot jump conditions, analysis is made to show the basic difference and interrelationship between acoustic and shock impedances of solids. The investigation also arrives at a new formulation for the general equation of shock and particle velocities in both binomial and polynomial forms. The nonlinear binomial should be of special interest to those who used to work on shock waves in solids.

1. INTRODUCTION. The Riemann invariant is well known in connection with the hyperbolic-type partial differential equations and with the transient analysis of nonlinear waves or supersonic flows. Thus, the method of characteristics can aid considerably the investigation of such relevant problems as shot propulsion or gun firing (internal ballistics), muzzle and recoilless-rifle blasts (transition or intermediate ballistics), and hypervelocity impact on armor (terminal ballistics). This paper is concerned with a basic analysis regarding the terminal ballistic effects in solids. For a given compression of solids, we can put its isentrope and shock adiabat in one-to-one correspondence by using the Mie-Grüneisen equation of state in coupled form. Such consideration turns out to yield a number of useful results and interrelationships between the acoustic and shock-wave properties of solids[1,2]. From yet another point of view we can determine a sound wave which corresponds to a shock wave with the same amount of compression. This calls for a transformation from the pv-plane to the pu-plane. Now each wave is associated with an impedance of its own just as it has its own pressure amplitude. Some results from this approach have been discussed in a recent paper[3], and further results are given in this paper for the higher-order effects and interrelationship between the two impedances. It may be noted that the Grüneisen parameter provides an interlink in the pv-plane[2] and that the Riemann invariant is the similar tool in the pu-plane[4]. From Riemann's invariant we can deduce the acoustic impedance and its higher-order effects. Likewise, we can evaluate the shock impedance using Hugoniot's momentum equation. Interrelations between the two impedances are best demonstrated by consideration of a third-order representation, which turns out to yield the generalized relation between the shock and particle velocities as has been used without proof in Reference 1. Now a complete derivation

$$(dp_s/du)_o = (dp_H/du)_o \tag{15'}$$

$$(d^2p_s/du^2)_o = (d^2p_H/du^2)_o \tag{16'}$$

$$(d^3p_s/du^3)_o = (d^3p_H/du^3)_o + 1/2(dp_H/du)_o^{-1}(d^2p_H/du^2)_o \times [1/2(d^2p_H/du^2)_o - G_o] \tag{17'}$$

Thus, the isentrope coincides initially with the shock adiabat in both pv- and pu-plane. The two adiabatic curves separate from each other on the third order as described by Eqs. (17) and (17'). These interrelations are very useful in this investigation. The following are all straightforward deduction.

From Eqs. (1)-(3), we may write

$$\begin{aligned} Z &= Z_o + Z_o'u + 1/2Z_o''u^2 + \cdots \\ &= (dp_s/du)_o + (d^2p_s/du^2)_o u + 1/2(d^3p_s/du^3)_o u^2 + \cdots \end{aligned} \tag{19}$$

Let us rearrange Eq. (4) to denote the shock impedance by $Y = U/v_o = p_H/u$ in analogy to Eqs. (1) and (19). Then we get

$$\begin{aligned} Y &= Y_o + Y_o'u + 1/2Y_o''u^2 + \cdots \\ &= (dp_H/du)_o + 1/2(d^2p_H/du^2)_o u + 1/6(d^3p_H/du^3)_o u^2 + \cdots \end{aligned} \tag{20}$$

It is interesting to note

$$Y_o = (dp_H/du)_o = (dp_s/du)_o = Z_o \tag{21}$$

$$Y_o' = 1/2(d^2p_H/du^2)_o = 1/2(d^2p_s/du^2)_o = 1/2Z_o' \tag{22}$$

$$\begin{aligned} Y_o'' &= 1/3(d^3p_H/du^3)_o = 1/3\{(d^3p_s/du^3)_o - 1/2(dp_s/du)_o^{-1} \\ &\quad (d^2p_s/du^2)_o \times [1/2(d^2p_s/du^2)_o - G_o]\} = \\ &\quad 1/3\,[Z_o'' - 1/2Z_o'(1/2Z_o' - G_o)/Z_o] \end{aligned} \tag{23}$$

by use of Eqs. (1)-(4), (15')-(17'), (19), and (20). Substituting Eqs. (21)-(23) and (19) in Eq. (20), we get

$$Y = Z - 1/2[Z_o'u+(2/3Z_o''+ \frac{1}{12} Z_o^{-1}Z_o'^2 - \frac{1}{6} G_oZ_o^{-1}Z_o')u^2] \quad (24)$$

which provides an interlink between the two compression impedances. Clearly, Eq. (24) is accurate to the third order of the adiabats. It is of particular interest to note that Eq. (24) is essentially the same as

$$U = a_o + 1/2Z_o'v_ou + 1/6[Z_o'' + 1/2Z_o'(G_o - 1/2Z_o')/Z_o]v_ou^2$$
$$= a_o + a_1u - a_2u^2 \quad (25)$$

with $U_o = a_o$ (or $Y_o = Z_o$). We shall consider Eq. (25) furthere in detail shortly.

3. WEAK NONLINEARITY. Thus far, we have expressed the impedances in polynomials. See Eqs. (19) and (20). It is still more illuminating to use the compact form:[5]

$$Z = Z_o(1 + p_s/B)^{N/2} \quad (26)$$

with constants

$$N = (K_o' + 1)^2/[K_o'(K_o'+1) - K_oK_o''] \quad (27)$$

$$B = (K_o' + 1)K_o/[K_o'(K_o' + 1) - K_oK_o'']. \quad (28)$$

Here K_o, K_o' and K_o'' are the initial values of $K = -vdp_s/dv$, $K' = dK/dp_s$ and $K'' = d^2K/dp_s^2$ at $p_s = 0$, respectively. From Eqs. (1) and (25)-(28), we get

$$a_o = v_oZ_o = (v_oK_o)^{1/2} \quad (29)$$

$$a_1 = 1/2v_oZ_o' = Nv_oZ_o^2/4B = (K_o' + 1)/4 \quad (30)$$

$$a_2 = (-v_o/6)[Z_o'' + 1/2Z_o'(G_o-1/2Z_o')/Z_o] = (-v_o/6)\,x$$
$$[1/2(N/B)^2 \times (1-N^{-1})Z_o^3 + (NZ_o/4B)(G_o-NZ_o^2/4B)]$$
$$= \frac{1}{12} (v_oK_o)^{-1/2}[\frac{1}{8} (K_o' + 1)(K_o'-4\gamma_o-7)-K_oK_o'']. \quad (31)$$

Now Eq. (31) may be rearranged as

$$-K_0K_0'' = 2a_1(\gamma_0-a_1+2) + 12a_0a_2 \tag{32}$$

which verifies the explicit results of Reference 1 using the Slater and Dugdale-MacDonald formulas. Here Eqs. (29)-(31) complete the formulation of Eq. (25) which was used without proof in our earlier paper.[1] It should be noted that the parabolic Eq. (25) may become useless well before it yields U<u. For this reason and by analogy with Eq. (26), let us consider the nonlinear binomial:

$$U = U_0(1 + u/\omega)^n \tag{33}$$

with $0<n\lesssim 1$. From Eqs. (25), (29)-(31), and (33), we get

$$n = a_1^2/(a_1^2 + 2a_0a_2)$$
$$= 3/4(K_0'+1)^2/[(K_0'+1)(K_0'-\gamma_0-1)-2K_0K_0''] \tag{34}$$

$$\omega = a_0a_1/(a_1^2 + 2a_0a_2)$$
$$= 3(K_0' + 1)(v_0K_0)^{1/2}/[(K_0'+1)(K_0'-\gamma_0-1)-2K_0K_0'']. \tag{35}$$

Now Eq. (33) is not only in closed form, but it is more general and more accurate than Eq. (25) provided that adequate data have been fitted in Eqs. (34) and (35) to determine n and ω. Thus, for $a_2 = a_1 = 0$ we get $n = 0$ and $U = U_0 = a_0$ (the acoustic approximation of very weak shocks). For $a_2 = 0$ only, we get $n = 1$, $\omega = a_0/a_1$, and $U = a_0 + a_1u$ (the widely-used, linear binomial). Elsewhere[4] we have shown this exactly by putting $N = 1$ in Eq. (26). Clearly, Eq. (33) can be expanded to yield the truncated form of Eq. (25) and

$$p_H(\varepsilon) = \varepsilon a_0^2 v_0^{-1} [(1-a_1\varepsilon)^2 + 2a_0a_2\varepsilon^2]^{-1} \tag{36}$$

as given in Reference 1 with $\varepsilon = 1 - v/v_0 = u/U$. If we combine Eqs. (4) and (33) with ε, we can determine a nonlinear adiabat $p_H = p_H(\varepsilon)$ which should be more accurate than the truncated form of Eq. (36).

4. CONCLUSION. In this investigation we consider the essential behavior and properties of shock waves in solids from a semi-analytical approach. Our new formulation of the velocity relation is not only simple in expression but also general, including such special cases as the binomial (linear) and polynomial (parabolic) forms which are being used widely in the literature.

REFERENCES

1. Y. K. Huang, "Shock-wave behavior and properties of solids", J. Appl. Phys. 42, 3212-3215 (1971).

2. Y. K. Huang, "Interrelationship between acoustic and shock-wave properties of solids", J. Appl. Phys. 42, 4084-4085 (1971).

3. Y. K. Huang, "Acoustic and shock impedances of solids at very high pressure", Proceedings Eleventh Annual Meeting of the Society of Engineering Science, ed. G. J. Dvorak (Duke University, 1974), pp. 88-89.

4. Y. K. Huang, "Note on shock compression of solids", J. Appl. Phys. 45, 2346-2347 (1974).

5. Y. K. Huang and C. Y. Chow, "The generalized compressibility equation of Tait for dense matter", J. Phys. D: Appl. Phys. 7, 2021-2023 (1974).

FREQUENCY DEPENDENT WAVE ARRIVAL TIME DELAYS IN DISPERSIVE AND NONDISPERSIVE MEDIA

J. R. Stabler, E. A. Baylot, and D. H. Cress
Mobility and Environmental Systems Laboratory
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. This paper presents the results of the application of a technique for estimating seismic wave arrival time delays in dispersive media. The technique addresses the physical situation consisting of two spatially separated sensors (seismic) implanted in a dispersive media for which the frequency-dependent time delays between the respective signatures are desired. The time delay at a given frequency is expressed in terms of the distance between the sensors and the frequency-dependent velocity of propagation. Initially, the phase differences between the two signatures are estimated. The accuracy of the estimated phase difference at each frequency is then associated with the coherence estimate. Phase differences for which the magnitude of the coherence is sufficiently large are used to estimate the time delays. Results of application of the technique are presented, and problem areas are discussed.

1. INTRODUCTION. Time delays between two measurements of wave motion in dispersive and nondispersive media play an important role in several fields of endeavor. Generally, time delays between two (or more) signatures are desired when the signatures are obtained from spatially separated sensors receiving energy from a common source as illustrated in Figure 1. Such physical situations may occur in such fields as oceanography (i.e. measured time delays between signatures of wave motion at two locations are used either to predict wave motion at a given location given the measured signature at another or to establish the degree of correlation between events at the two locations), seismology (i.e. measured time delays are inserted into bearing location algorithms to locate earth tremors and explosions), or acoustics (i.e. based upon measured time delays, velocities of propagation of energy through various media are determined).

The purpose of this paper is to review the conventional approach for making frequency dependent estimates of time delays* using the fast Fourier transform (FFT),** to illustrate the results of application of that approach to particular types of signatures (seismic signatures of military vehicles), to point out its limitations and to discuss a method for avoiding some of those limitations.

---

* R. K. Otnes and L. Enochson, Digital Time Series Analysis, John Wiley, New York, 1972.

** J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," Mathematics of Computation, Vol 19, 1965, p 297.

2. THEORETICAL CONSIDERATIONS. The physical situation addressed in this analysis consists of two sensors spatially separated by some medium (dispersive or nondispersive). If the energy source is at a sufficient distance from the two sensors, the incoming waves can be described by plane wave fronts (Figure 1). The line between these sensors is chosen to define the z axis. The energy is assumed to be propagating in the positive z direction (away from the source). The particle displacement for the selected mode of propagation is assumed to be in one direction, although that direction is arbitrary (i.e. depending upon whether or not the vibration is compressional, transverse, etc.).* The assumption of a single mode of propagation is employed to avoid the ambiguity that may occur in estimating time delays when several vibrational modes having different propagational velocities are present, hence leading to different time delays for each mode of vibration.

For a particular point along the z axis, and for a particular frequency, the displacement of particles can be described by the conventional expression

$$x(z,t) = A \exp j(2\pi ft - kz)$$

where

x = displacement
t = time
A = amplitude
f = frequency
k = wave number = 2π/wavelength

In practice it is generally the rate of change of x with respect to time that is measured, or

$$\dot{x}(z,t) = 2\pi fA \exp j(2\pi ft - kz + \frac{\pi}{2}) \quad .$$

If we consider the effective change in x at these two sensore (i.e. sensor 1 at $z = 0$ and sensor 2 at $z = d$), the wave form of these two sensors becomes

$$\dot{x}_1(o,t) = 2\pi fA \exp j(2\pi ft + \frac{\pi}{2})$$

$$\dot{x}_2(d,t) = 2\pi fA \exp j(2\pi ft - kd + \frac{\pi}{2}).$$

The difference between the above wave forms is due to a phase difference $\phi = -kd$.

The wave number k can be expressed in terms of the velocity ot propagation v and the frequency as $k = 2\pi f/v$. Therefore,

---

* Actually, the assumption ot motion along a single direction can be relaxed to include motion having only one independent variable such as for Rayleigh waves where the motions in orthogonal directions are dependent on one another.

$$\phi = -2\pi f\left(\frac{d}{v}\right)$$

The time delay between signatures received at the sensors is the quantity d/v and is therefore given by

$$t = \frac{\phi}{2\pi f}$$

In the actual physical situations, the energy is spread over bands encompassing a broad range of frequencies. When a broad region of frequencies is present in the signatures, a plot of phase difference versus frequency can be generated. For nondispersive media the velocity is constant so that the phase shift increases linearly with frequency as indicated in Figure 2a. For dispersive media the phase shift may become a complex function of frequency as indicated in Figure 2b. The time delay between two signatures is directly related to the phase difference ϕ as indicated in equation 1. Therefore, emphasis in the remainder of this paper is upon determination of ϕ, since it is the phase shift that is directly obtainable from the conventional frequency-by-frequency comparison of two signatures.

With the advent of the fast Fourier transform (FFT), the complete spectrum can be computed in reasonable computational time. It is now practical to compute the phase difference for each discrete frequency band resolved in the frequency domain. The conventional approach consists of the following steps:

a. The time domain signatures received at two sensors, x and y, shall be denoted by $\dot{x}(t)$ and $\dot{y}(t)$. The frequency domain representation of $\dot{x}(t)$ and $\dot{y}(t)$ is obtained by application of the discrete Fourier transform to obtain

$$\tilde{X}_k = \Delta t \sum_{m=0}^{n-1} \dot{x}_m \exp\left(-j2\pi f_k m\Delta t\right)$$

$$\tilde{Y}_k = \Delta t \sum_{m=0}^{n-1} \dot{y}_m \exp\left(-j2\pi f_k m\Delta t\right)$$

where Δt is the sampling period, $\dot{x}_m$ and $\dot{y}_m$ are $\dot{x}(m\Delta t)$ and $\dot{y}(m\Delta t)$, respectively, and X_k and Y_k are complex variables with real and imaginary parts. The real part of $\tilde{X}_k$ (or $\tilde{Y}_k$) denotes that part of $\dot{x}(t)$ [or $\dot{y}(t)$] that can be associated with the cosine terms in Fourier expansion while the imaginary part can be associated with the sine terms in the Fourier expansion. The "tildes" over $\tilde{X}_k$ and $\tilde{Y}_k$ indicate that $\tilde{X}_k$ and $\tilde{Y}_k$ are "raw"

estimates of the frequency domain signatures without any averaging over adjacent frequency values or sequential time estimates.

b. The cross-spectral density estimate, G_{xyk}, is obtained from $\tilde{X}_k$ and $\tilde{Y}_k$ using the defining expression

$$\tilde{G}_{xyk} = \tilde{X}_k^* \tilde{Y}_k$$

where the asterisk denotes complex conjugate. The resulting expression for G_{xyk} can be separated into real and imagery parts in the form

$$\tilde{G}_{xyk} = \tilde{C}_{xyk} - j \tilde{Q}_{xyk}$$

where $\tilde{C}_{xyk}$ is referred to as the cospectrum estimate and $\tilde{Q}_{xyk}$ is the quadrature spectrum estimate. The power spectrum, or autospectrum, is estimated for both $x(t)$ and $y(t)$ from the expression

$$G_{xk} = X_k^* X_k$$

and

$$G_{yk} = Y_k^* Y_k .$$

c. The raw cross-spectral estimates can be smoothed over M adjacent estimates using the expressions

$$\hat{G}_{xk} = \frac{1}{M} \sum_{j=1}^{M} G_{x(k+j)}$$

$$\hat{G}_{yk} = \frac{1}{M} \sum_{j=1}^{M} G_{y(k+j)}$$

$$\hat{G}_{xyk} = \frac{1}{M} \sum_{j=1}^{M} G_{y(k+j)}$$

$$= \hat{C}_{xyk} - j \hat{Q}_{xyk}$$

The value of M selected for the smoothing is dependent upon the physical properties of the medium through which the energy propagates, the desired accuracy of the estimates, and the desired frequency resolution. As the value of M is increased, the bandwidth of the resolved frequencies is increased, and the accuracy of the estimates are improved. However, the bandwidth of the resolved frequencies is restricted by the degree of dispersion (i.e. the dependence of velocity on frequency). This is because improvement in the accuracy of estimates with increased values of M is dependent upon the assumption that the velocity of propagation of energy is reasonably constant in the bandwidth of resolution of frequencies. Therefore, M is restricted to values for which this assumption is true.

d. The phase difference, ϕ_{xyk}, between time series x(t) and y(t) is estimated from the expression

$$\hat{\phi}_{xyk} \quad \text{arc tan } (\hat{Q}_{xyk}/\hat{C}_{xyk}) .$$

An additional parameter having useful properties for the application of the computational technique to time-delay estimates,* or equivalently, estimates of phase difference, is the squared coherence, γ^2_{xy}. This parameter is defined in terms of the cross spectrum $G_{xy}(f)$ and the autospectra $G_x(f)$ and $G_y(f)$ as

$$\gamma^2_{xy}(f) = \frac{|G_{xy}(f)|^2}{G_x(f)\ G_y(f)}$$

The coherence estimate can be expressed in terms of the previously obtained estimates of power and cross spectra as

$$\gamma^2_{xy} = \frac{|\hat{G}_{xyk}|^2}{\hat{G}_{xk}\ \hat{G}_{yk}} = \frac{\hat{C}^2_{xyk} + \hat{Q}^2_{xyk}}{\hat{G}_{xk}\ \hat{G}_{yk}}$$

3. APPLICATION. Application of the conventional approach described previously for computation of the phase difference between two signatures may produce inconsistent results. For example, the calculated

---

* B. V. Hammon and E. J. Hannan, "Spectral Estimations of Time Delays for Dispersive and Non-Dispersive Systems," Journal of Applied Statistics, Vol 23, 1974, pp 134-142.

phase difference between two signatures can be so irratic in some frequency bands that it is impossible to identify which estimates, if any, are reliable. As discussed by B. V. Hammon and E. J. Hannan,* the coherence parameter (γ^2_{xy}) can be directly related to the accuracy of the cross-spectral estimate (hence phase difference) at a particular frequency. Because the accuracy of the phase-difference estimate improves rapidly as the coherence increases toward unity, the coherence parameter can be used to determine what frequencies can be used to provide the best estimates of phase difference. Results of applying the conventional approach to phase-difference estimates, ignoring coherence, are presented below. Problem areas are identified. The coherence parameter is then included in the analysis and the resulting contribution of coherence to identifying reliable phase difference estimates is discussed.

Application excluding coherence. The first calculation of phase difference considered is the result of analysis of seismic signatures of an M113 APC (armored personnel carrier) collected with two geophones spaced 6 m apart. The phase plot (Figure 3a) suggests that there is a continuous functional dependence between the frequency and phase difference in the approximate frequency bands 5-25 Hz and 50-210 Hz. However, scattering in the phase-difference estimates is clearly evident between 25-50 Hz and at frequencies greater than 210 Hz with some scattering in all bands. Observation of the power spectrum (Figure 3b) shows that power is, on the average, somewhat larger in the 50-210 Hz band than in the 25-50 Hz and 210-250 Hz bands and considerably larger in the 0-25 Hz region. It can be postulated that the contributions of random noise in the instrumentation and background seismic noise combine to mask the contributions of the signature of the M113 APC in low-energy spectral regions. The resulting estimate of phase difference in the "lower energy" bands appear to be random.

Often it is desirable for the investigator to determine phase differences from two or more energy sources simultaneously. Such is illustrated in Figure 4; two seismic sources are present, and M151 jeep and an M35 truck. An ambiguity then arises as to what the "correct" time delay is. For instance, the effective distances between the geophones for each of the two sources are different so that the resulting time delays in the reception of the respective energy contributions of the two sources are different. This ambiguity results in scattering of the estimates of phase difference in the frequency regions where the energy from the respective sources overlap. Estimates of phase difference for the physical situation illustrated in Figure 4 are presented in Figure 5a. The estimates are noticeably scattered. Inspection of the power spectrum (Figure 5b) shows that the energy is again primarily concentrated below 25 Hz. However, several strong lines appear in the spectrum (for example at approximately 80, 90, and 170 Hz) suggesting that, at least in these narrow frequency bands, one source could be dominating over another (i.e., it would be unlikely that the two sources

*Ibid., p 5.

would generate peaks of energy in the same narrow frequency bands). It should be possible to identify reliable phase-difference estimates (hence time delays) for these narrow frequency bands. However, reliable estimates may exist in bands not readily identifiable in the power spectrum.

Application including coherence. In view of the difficulties encountered in estimating phase differences between signatures generated by multiple sources or containing random noise, some criterion is desirable for selecting potentially accurate phase-difference estimates. As discussed by Hammon and Hannan,* the coherence parameter provides an estimate of the accuracy of the phase-difference estimates. The coherence spectra for the two cases discussed previously are presented below and the phase differences are discussed in view of the associated coherence spectra.

The coherence spectrum for the single energy source (M113 APC) is presented in Figure 6a. The average value of the coherence is relatively low in the frequency region between 25 and 50 Hz and for frequencies greater than 210 Hz. There are large variations in the coherence in the 50-210 Hz region. If the phase difference is plotted only for values of coherence greater than 0.5, only the phase difference appearing in Figure 6b appears. Comparison of Figure 6b with the previous plot of phase difference (Figure 3a) reveals that many estimates have been eliminated, particularly those having no correlation to adjacent estimates. The coherence criterion suppresses the random components of the estimates of phase difference.

The coherence spectrum for the multiple energy sources is presented in Figure 7a. The coherence is generally lower than that of the single-energy source. The phase difference is plotted for values of coherence greater than 0.5 in Figure 7b. Comparison of Figure 7b with Figure 5a reveals that the number of phase differences that satisfy the coherence criterion is significantly less than the number of available estimates. "Clusters" of three or four estimates satisfying the coherence criterion occur at approximately 80, 90, and 180 Hz, corresponding to lines in the power spectrum. However, other phase differences also satisfy the criterion, indicating that inspection of the power spectrum alone does not necessarily allow one to identify all of the reliable estimates of phase differences.

Summary. Two sets of data have been analyzed to obtain the phase difference between signatures for subsequent determination of time delays. Random instrument noise and the interaction between the signatures of multiple energy sources have been hypothesized as potential causes for scattering of the estimates of phase differences. Those estimates of phase difference useful for estimating time delays have been identified by applying a criterion on coherence to the estimates of phase difference. The coherence estimate is easily implemented in the conventional approach to obtaining estimates of phase difference, since the supporting parameters (i.e. X_k , Y_k , G_{xyk} , etc.) must be calculated in order to obtain the phase difference.

*Ibid., p 5.

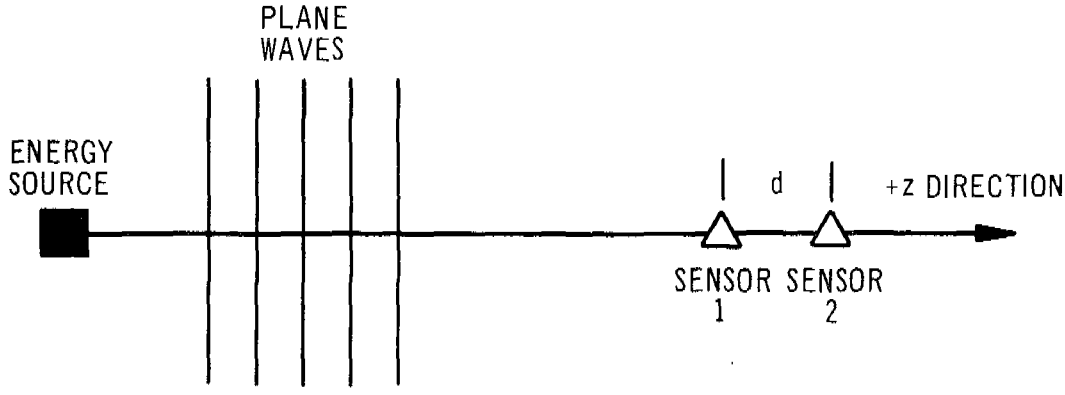
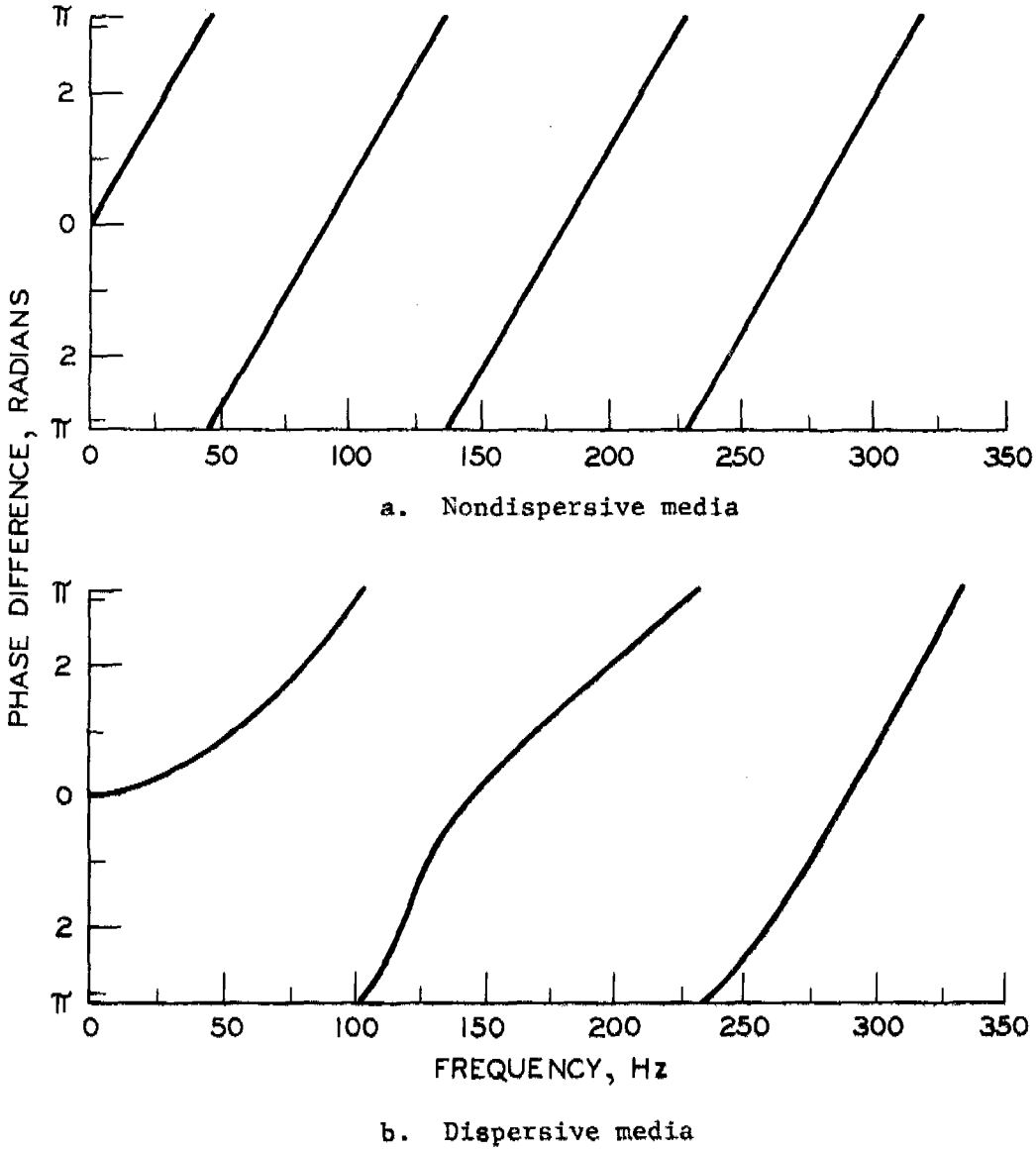


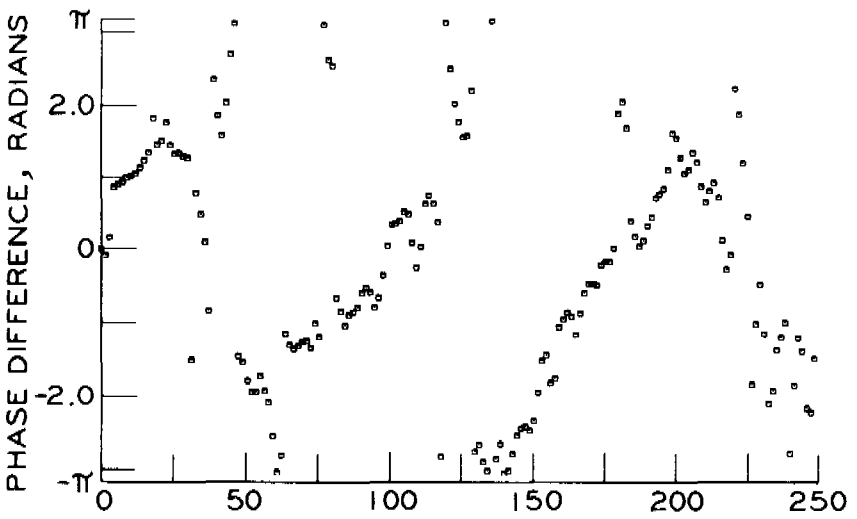
Figure 1. Geometric relation between sensors and incoming plane waves



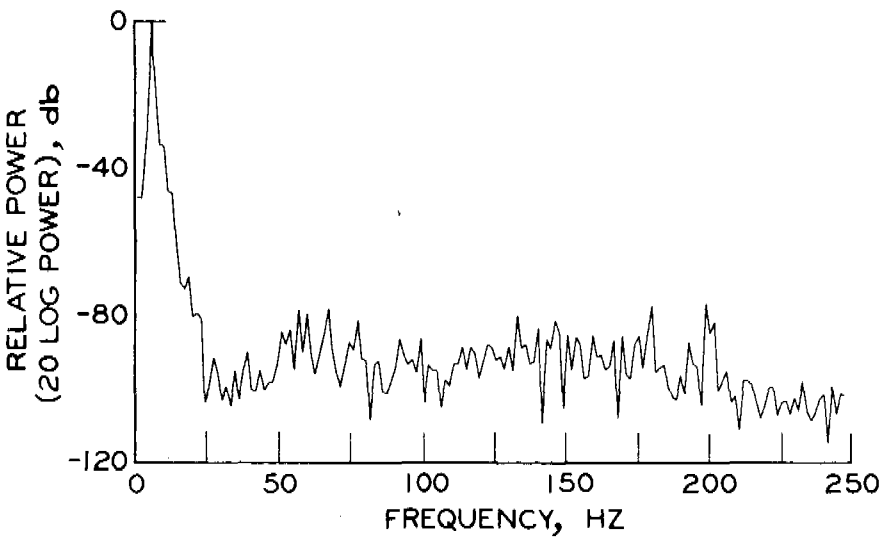
a. Nondispersive media

b. Dispersive media

Figure 2. Phase differences for nondispersive and dispersive media



a. Phase difference



b. Power spectrum

Figure 3. Phase difference and power spectrum for single source (M113 APC)

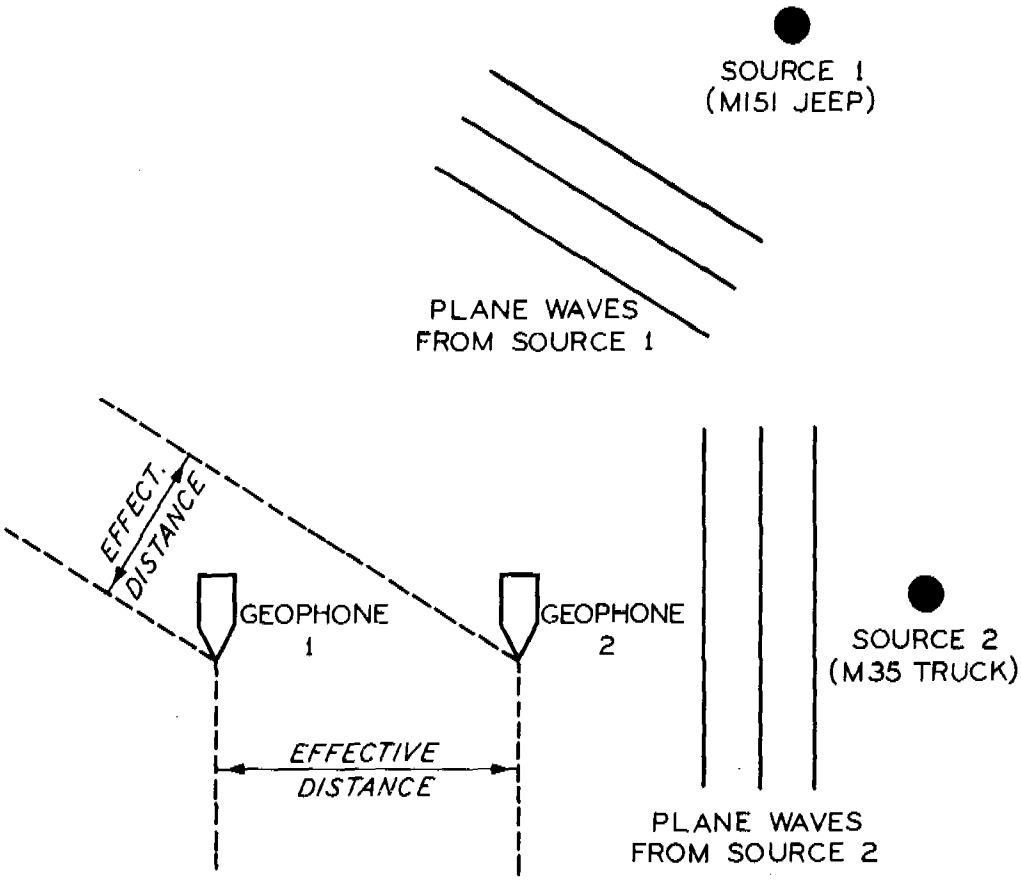
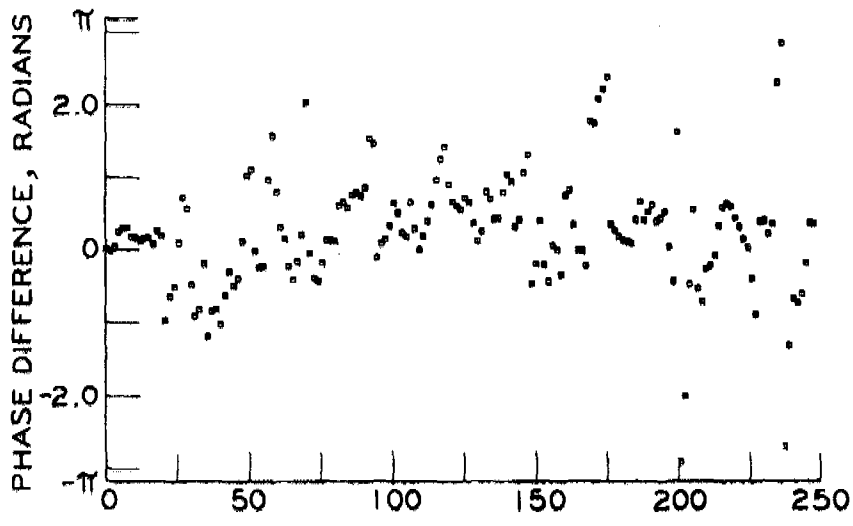
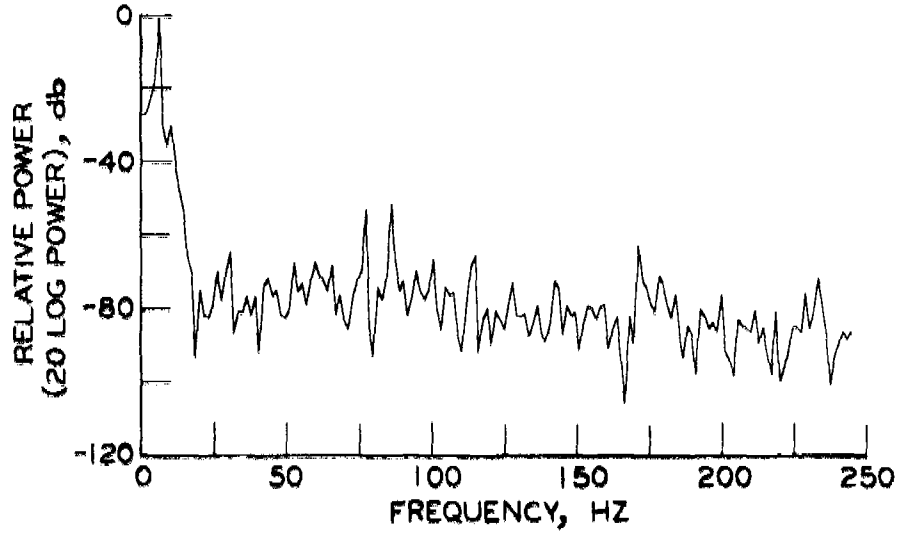


Figure 4. Physical description of multiple sources (M151 jeep and M35 truck) study

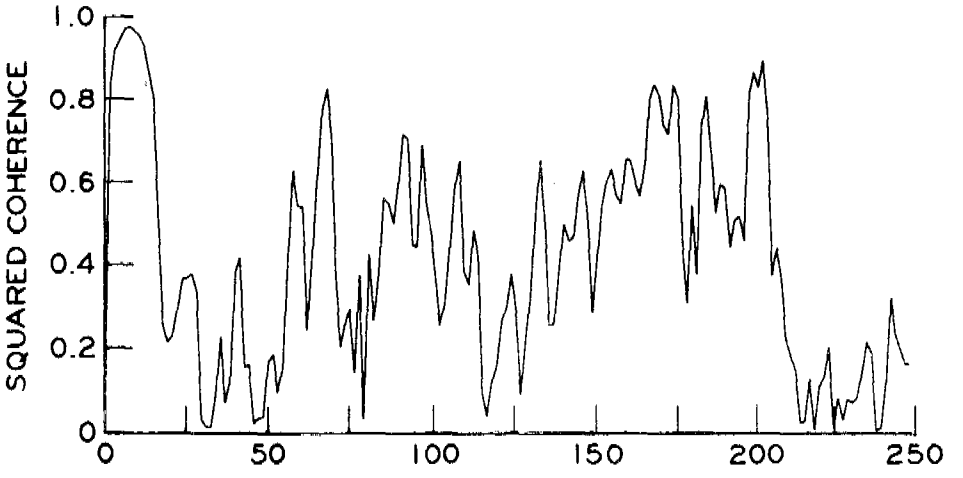


a. Phase difference

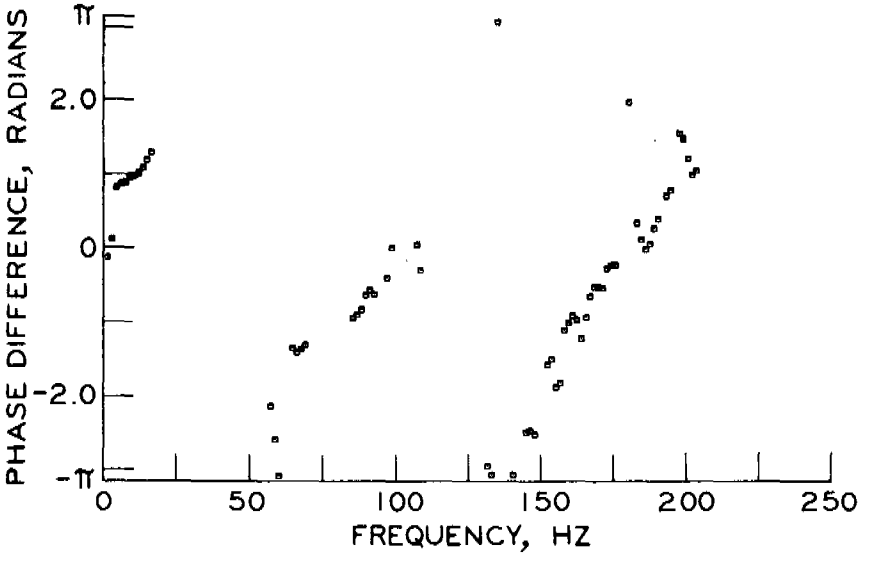


b. Power spectrum

Figure 5. Phase difference and power spectrum for multiple sources (M151 jeep and M35 truck)

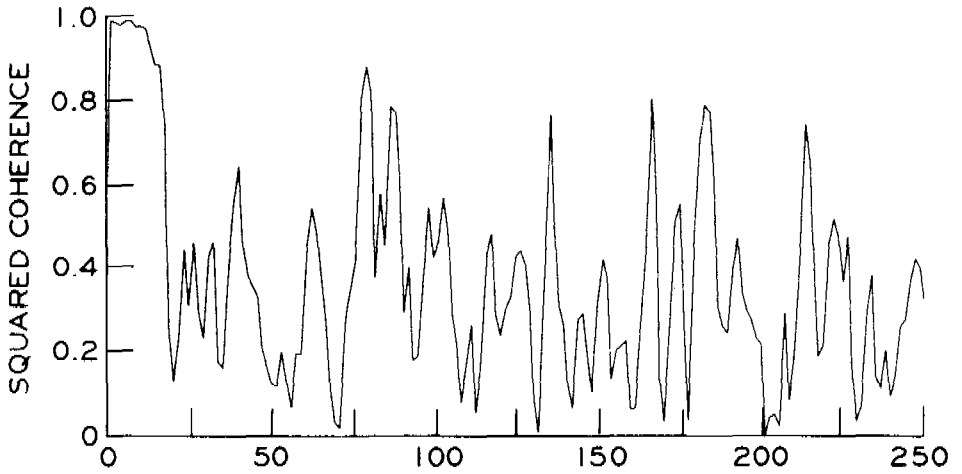


a. Coherence spectrum

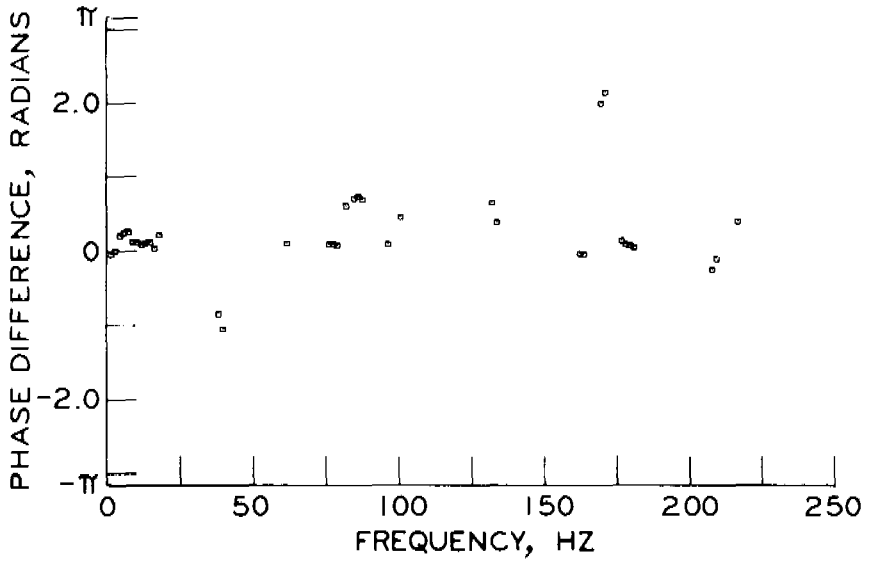


b. Phase difference

Figure 6. Coherence spectrum and phase difference for single source (M113 APC)



a. Coherence spectrum



b. Phase difference

Figure 7. Coherence spectrum and phase difference for multiple sources (M151 jeep and M35 truck) with square coherence values greater than 0.5

THE METHOD OF PARABOLIC SUBSTITUTION FOR HIGH SUBSONIC FLOW

Klaus Oswatitsch

Technische Hochschule Wien
Director of the Institute of Theoretical Gasdynamics
DFVLR, Aachen, West Germany

Robert E. Singleton

U. S. Army Research Office
Research Triangle Park, N. C. 27709

ABSTRACT. A new relaxation method is formulated by substituting artificial time dependence for true time dependence in the Euler equations of fluid mechanics. The previously hyperbolic equations are thereby converted into parabolic form, and the convergence question is studied analytically within the vehicle of linearized small disturbance theory. The fully nonlinear equations are solved numerically using this new relaxation technique for the case of a NACA 0012 airfoil at 2° angle of attack. The method is seen to work quite well and warrants further extension into the transonic range.

1. INTRODUCTION. The problem of determining the flow field past an arbitrary, two-dimensional body immersed in an unbounded compressible gas has been the subject of countless investigations over the last half-century. Most of these analytical methods gave reasonable results until the free stream Mach number increased to the point that a finite supersonic zone usually terminated by a shock wave appeared on the body. This transonic flow regime has only during the last few years yielded to the attempts of many researchers to develop solution techniques general enough to be utilized as design tools. These successful techniques are primarily numerical relaxation methods applied to the entire flow field equations.

In the current work described here, the fully nonlinear equation is solved utilizing transformation technique for handling boundary conditions. However, the relaxation process itself is quite novel and the physical motivation described in deriving the relaxation method presented here makes it quite possible to devise a relaxation process which converges much faster than any technique currently being used. In the sections to follow, the derivation of the relaxation method will be presented followed by a discussion of the question of convergence. The application of this technique to subsonic flow is presented in Sections 4 and 5, and results are discussed in Section 6.

2. MOTIVATION AND DERIVATION OF METHOD. The selection of any particular iterative method of successive approximation for finding the solution of a set of nonlinear, algebraic equations is determined firstly, by the conditions for convergence of the method and secondly, by the speed with which the method converges. The concept of utilizing the fully unsteady equations of motion to obtain steady-state flow fields is particularly intriguing from the point of view of a relaxation scheme selection since one can argue from physical experiments that initially unsteady transonic flow fields with steady boundary conditions do indeed settle down to steady-state conditions. Hence a relaxation process which faithfully describes this true transient behavior should always converge to the desired steady-state solutions. The equations to be solved are of hyperbolic type and hence the flow pattern at any finite instant of time is composed of a complex pattern of rarefaction and compression waves. The numerical description of this complex process together with the boundary conditions, even though guaranteed to converge if the difference scheme is stable, requires an enormous amount of real computer time, i.e., the speed of convergence is slow.

To obtain a relaxation process which converges more rapidly, perhaps some other artificial time behavior should be utilized rather than the true time behavior of the hyperbolic Euler equations for unsteady flow. The motivation for this concept is that hyperbolic systems approach their asymptotic steady states through a series of expansions and compressions, i.e., through a wave system. Now if one substitutes for this true time behavior, an artificial time behavior which yields a parabolic set of equations, then the resulting solution should settle down to its asymptotic steady state through a process of diffusion. The attempt here is to draw an analogy between heat diffusion processes which are governed by parabolic type equations and the unsteady fluid flow equations which have been manipulated, albeit artificially, to be parabolic equations. Carrying the analogy further, since heat diffusion processes have been observed to reach their asymptotic steady state rather quickly, exponentially fast in some cases, it then might be expected that the fluid flow should also approach its steady state more quickly through an artificial time-dependent process which is of parabolic character. That is to say, a hyperbolic or wave-type set of equations has been traded for a parabolic or diffusion-type set of equations. It must be remembered, however, that the quantity called time in the parabolic equations can no longer be interpreted as a physically meaningful quantity but takes on the interpretation of an artificial parameter which gauges the nearness to an asymptotic steady-state solution. The desired solution is obtained when dependent flow variables no longer depend on this artificial parameter. At this point, it is reasonable to ask if these parabolic equations converge to some steady-state solution and if the solution so obtained is the correct one. This question will be discussed in Section 3. To specifically

define this new relaxation concept, let ρ = density, q = velocity vector, p = pressure, Q = heat per unit mass and time, i = enthalpy, R = gas constant. Then the governing equations for unsteady flow of an inviscid perfect gas past an obstacle are

$$(1) \qquad \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho q) = 0$$

$$(2) \qquad \frac{\partial q}{\partial t} + (q \cdot \nabla) q = - \frac{1}{\rho} \nabla p$$

$$(3) \qquad \frac{\partial}{\partial t}\left(i + \frac{1}{2} q^2\right) + (q \cdot \nabla)\left(i + \frac{1}{2} q^2\right) = \frac{1}{\rho} \frac{\partial p}{\partial t} + Q$$

$$(4) \qquad p = \rho R T.$$

The boundary conditions for an obstacle immersed in an unbounded fluid are

$$(5) \qquad q_n = 0, \quad q = q_\infty$$

where q_n is the velocity normal to body surface and q_∞ is the velocity vector at infinity. In addition, for flows with circulation, the Kutta condition,

$$q_{T.E.} = 0,$$

where $q_{T.E.}$ is the velocity at the trailing edge, must be satisfied. For free-stream Mach number ranging up to transonic values, the flow field can be assumed to be both isentropic and irrotational at least to second order since shock waves, if present, are weak and therefore produce entropy jumps of third order. Eq. (3) is consequently replaced by

$$(6) \qquad \frac{p}{\rho^\gamma} = \text{CONSTANT}$$

Now defining i_0 = stagnation enthalpy, a_∞ = free-stream speed of sound, $h = i + q^2/2 - i_0$, the conservation of momentum and mass becomes, respectively,

(7) $$\frac{\partial q}{\partial t} + \nabla h = 0$$

(8) $$\frac{\partial q^2}{\partial t} - \frac{\partial h}{\partial t} - (\gamma - 1)(\nabla \cdot q) h = a_\infty^2 \nabla \cdot q - (q \cdot \nabla) q^2/2 - \frac{(\gamma - 1)}{2}\left(q^2 - q_\infty^2\right)(\nabla \cdot q)$$

The expression on the left side of Eq. (8) lumps all the time derivatives into one term, which can be written

(9) $$-2(q \cdot \nabla)h - \frac{\partial h}{\partial t} - (\gamma - 1)(\nabla \cdot q)\, h = -J(h)$$

Now if distance, velocity, q, time, t, and h are interpreted as having been scaled with respect to some characteristic length, L, free-stream velocity, q_∞, characteristic time, L/q_∞, and q_∞^2, respectively, and defining M_∞ = free-stream Mach number,

$$J(h) = \frac{q_\infty^3}{L M_\infty^2} F(h)$$

then Eq. (8) becomes

(10) $$\frac{1}{M_\infty^2} \nabla \cdot q - (q \cdot \nabla) \frac{1}{2} q^2 - \frac{\gamma - 1}{2}\left[q^2 - 1\right] \nabla \cdot q + \frac{1}{M_\infty^2} F(h) = 0 .$$

For irrotational flow, $q = \nabla \Phi$, and then from Eq. (7)

$$h = -\partial \Phi / \partial t .$$

Introducing the potential function in Eq. (10) gives

$$(10a)\qquad \frac{1}{M_\infty^2}\nabla^2\Phi - (\nabla\Phi\cdot\nabla)\frac{\nabla\Phi\cdot\nabla\Phi}{2} - \frac{\gamma-1}{2}\left(\nabla\Phi\cdot\nabla\Phi - 1\right)\nabla^2\Phi + \frac{1}{M_\infty^2}F(-\Phi_t) = 0 \ .$$

The method of parabolic substitution is to substitute a simple function for $F(-\Phi_t)/M_\infty^2$, termed $S(\Phi_t)$, which yields a parabolic equation and which, since steady-state solutions are being sought, satisfies the condition $S(0) = 0$.

Within these two restrictions the choice for $S(\Phi_t)$ is arbitrary. Of course having made a choice for S, the question of convergence as time increases of the resulting artificially unsteady flow to steady-state conditions arises.

3. THE CONVERGENCE QUESTION

A simple choice for S which satisfies the above two requirements is

$$(11)\qquad S(\Phi_t) = -\Phi_t \ .$$

To study the convergence of the resulting unsteady solution to the required steady-state values as time increases, a small disturbance analysis is made for symmetric flow past a thin two-dimensional obstacle. This problem is stated as

$$(12)\qquad \left(1 - M_\infty^2\right)\varphi_{xx} + \varphi_{yy} = \varphi_t$$

$$(13)\qquad \begin{cases} \left(\varphi_y\right)_{y=0^-} = Y(x)\,f(t) & (|x| < 2,\ t > 0) \\ \varphi_x,\ \varphi_y \longrightarrow 0 & \text{as } \sqrt{x^2 + y^2} \longrightarrow \infty \\ \left(\varphi_x\right)_{t=0} = \left(\varphi_y\right)_{t=0} = \left(\varphi_t\right)_{t=0} = 0 \end{cases}$$

where ϕ = perturbation potential, $Y(x)$ = surface slope of obstacle and

$$f(t) = \begin{cases} 0 \text{ for } t < 0, \\ 1 \text{ for } t > 0. \end{cases}$$

The above problem is linear and exact closed form expressions can be obtained for the potential function. When asymptotic expansions are made for large t, the solution yields

$$\phi_{x,y} = (\phi_{xy})_{\text{steady}} + 0(\frac{1}{t})$$

and therefore as $t \to \infty$, the approximate solution does converge to the correct steady-state solution.

To investigate another choice for $S(\phi_t)$, let

$$S(\Phi_t) = \frac{1}{4}\Phi_{tt} + \Phi_{yt} \tag{14}$$

which also satisfies the two requirements on S. The problem to be solved is the same as before except Eq. (12) is replaced by

$$\left(1 - M_\infty^2\right)\varphi_{xx} + \varphi_{yy} = -\frac{1}{4}\varphi_{tt} - \varphi_{yt} \tag{15}$$

The solution of this equation subject to conditions (13) can also be found analytically and expanded for large t to give

$$\phi_x = (\phi_x)_{\text{steady}} + 0(\frac{1}{t^2})$$

$$\phi_y = (\phi_y)_{\text{steady}} \qquad t > y/2$$

and therefore as $t \to \infty$, the approximate solution does converge to the correct steady-state solution. Since it is of interest to apply this new relaxation procedure to transonic flow eventually, the convergence question can also be studied within the vehicle of linearized transonic small disturbance theory together with the substitution of Eq. (11). Hence the equation is changed from Eq. (12) to

$$\left(1 - M_\infty^2\right)\varphi_{xx} + \varphi_{yy} - K\varphi_x = \varphi_t$$

where K is a positive constant representing acceleration at some point on the airfoil. Again, asymptotic expansions of the analytic solution are easily found to be

$$\phi_x = (\phi_x)_{\text{steady}} + 0(\frac{1}{t}\, e^{-at})$$

$$\phi_y = (\phi_y)_{\text{steady}} + 0(\frac{1}{t^2}\, e^{-at})$$

where a is a constant and hence the approximate solution is seen to converge to the correct steady-state solution. To examine this question of convergence for the actual nonlinear equation of interest requires a complete numerical analysis using high-speed computers.

4. PREPARATION FOR NUMERICAL ANALYSIS

The problem to be discussed in the remaining part of this paper is the numerical solution of Eq. (10a) with the substitution (11) for a high-speed flow past an arbitrary two-dimensional body. To enable an accurate numerical description of the boundary conditions a conformal transformation is made from the physical flow plane into the interior of a unit circle, hereafter called the computation plane. In doing thusly, the arbitrary body surface is transformed onto the unit circle in the computation plane and infinity in the physical plane is transformed into the center of the circle in the computation plane. A cylindrical polar coordinate system r, Θ is utilized in the computational plane so that r = 1 corresponds to the body surface and r = 0 corresponds to infinity. As an extra advantage, this mapping procedure distributes mesh points more densely in those regions in the computation plane where flow acceleration is the greatest, thus allowing for better flow resolution in those regions where variables are changing the most. For some body contours, such as Joukowsky or Kármán-Trefftz airfoils, this transformation can be computed analytically, however, for arbitrary bodies the method of D. Catherall, D. N. Foster and C. C. L. Sells [1] for calculating the transformation has been faithfully followed in this report.

In the general case of flows with circulation, the potential function has a dipole singularity at r = 0. Hence by comparison with incompressible flow, it is seen that if Φ is expanded about r = 0 in the computation plane, the form has to be

$$\Phi \sim \frac{\cos(\alpha + \alpha_1 + \theta)}{r} + L_1(\theta) + O(r) \tag{16}$$

where α = angle of attack with respect to x-axis (positive counter clockwise), α_1 = flow direction with respect to x-axis for zero lift in incompressible flow (positive clockwise), $L_1(\Theta)$ = unknown function of Θ.

Hence, to determine $L_1(\Theta)$, the transformed equation for Φ is expanded about r = 0 and the highest two orders are retained. The result is that

$$L_1(\theta) = \frac{E}{\sqrt{1 - M_\infty^2}} \tan^{-1}\left[\sqrt{1 - M_\infty^2}\, \tan(\alpha + \alpha_1 + \theta)\right]$$

where E is a constant to be determined from the Kutta condition at the trailing edge and

$$\tan^{-1}\left[\sqrt{1 - M_\infty^2}\,\tan(\alpha + \alpha_1 + \theta)\right]$$

is determined to have the same quadrant as $\alpha + \alpha_1 + \Theta$.

A disturbance potential, χ, is now introduced by the definition

$$\chi = \Phi(r,\theta,t) - \left[\frac{\cos(\alpha + \alpha_1 + \theta)}{r} + E(\theta + \alpha + \alpha_1) + r\cos(\alpha + \alpha_1 + \theta)\right] . \tag{17}$$

In terms of χ, the boundary conditions at r = 1 and r = 0 are

$$\chi_r(1,\theta,t) = 0 \tag{18}$$

$$\chi(0,\theta,t) = L_1(\theta) - E(\theta + \alpha + \alpha_1) . \tag{19}$$

By initially selecting

$$E = 2\sin(\alpha + \alpha_1)$$

and

$$\Phi(r,\Theta,0) = \Phi_{\text{incompr}}$$

then

$$\chi(r,\Theta,0) = 0. \tag{20}$$

As previously pointed out, E is determined by the Kutta condition at the airfoil trailing edge located at r = 1, Θ = 0. This condition requires

$$\left[\chi_\Theta\right]_{r=1,\Theta=0} = -E + 2\sin(\alpha + \alpha_1)$$

or

$$E(t) = -\chi_\theta(1,0,t) + 2\sin(\alpha + \alpha_1) \tag{21}$$

and hence E can be updated at each time level as the solution proceeds. The initial and boundary conditions have now been put in a form easily expressed by numerical techniques. However, since a potential function analysis has been used, there will, in general, be a branch cut in the computation plane defined by $\theta = 0$, $0 \leq r \leq 1$. Hence, when computing θ-derivatives along $\theta = 0$, central differences cannot be used but must be supplanted with second order accurate forward or backward differences along the rays $\theta = 0$, $\theta = 2\pi$, respectively. Since pressure must be continuous across this branch cut the velocity components on $\theta = 0$ must be the same as they are on $\theta = 2\pi$ for a given r. Hence along the branch cut one can write $2N_r + 2$ equations, where N_r is the number of mesh points in the radial direction excluding $r = 0$, for $2N_r + 2$ values of χ. This system of linear algebraic equations can be straightforwardly solved by method of elimination to establish the values of χ along the ray $\theta = 0$ and the ray $\theta = 2\pi$ in terms of neighboring χ values. The two rays, $\theta = 0$ and $\theta = 2\pi$ now are used as two additional boundaries to the numerical scheme.

The appropriate equation which now must be solved at the interior mesh points, $0 < r < 1$ and $0 < \theta < 2\pi$, is the transformed equation (10a) with the substitution of (11) for the artificial time dependence and the introduction of χ through Equation (17). This complicated equation takes the form

$$\mathcal{B}^2\left(r^2\chi_{rr} + \chi_{\theta\theta} + r\chi_r\right) + M_\infty^2\left(A\chi_{rr} + B\chi_{r\theta} + C\chi_{\theta\theta} + D\right) = \chi_t \tag{22}$$

where $\mathcal{B}$ = transformation modulus and is a function of r, θ,

$$A = a_1 + (a_2 + a_3E)E + (a_4 + a_5\chi_r)\chi_r + (a_6 + a_7E + a_8\chi_\theta)\chi_\theta,$$

$$B = (b_1 + b_2\chi_r)(b_3 + E + \chi_\theta),$$

$$C = c_1 + c_2(E + \chi_\theta) + c_3(E + \chi_\theta)^2 + (c_4 + c_5\chi_r)\chi_r,$$

$$D = (d_1 + d_2E + d_3E^2)\chi_r + (d_4 + d_5E + d_6E^2)\chi_\theta + (d_7 + d_8E)\chi_r^2 + (d_9 + d_{10}E)\chi_\theta^2 + (d_{11} + d_{12}E)\chi_r\chi_\theta + d_{13}\chi_r^2\chi_\theta + d_{14}\chi_\theta^2\chi_r + d_{15}\chi_r^3 + d_{16}(\chi_\theta^3 + E^3) + d_{17} + d_{18}E + d_{19}E^2,$$

and a_i, b_i, c_i and d_i are functions of r, Θ as determined by the computed conformal transformation.

5. NUMERICAL METHOD.

Eq. (22) is now differenced in the classic manner for parabolic equations. Second order accurate centered differences are used for spatial derivatives and the first order accurate Euler difference is used for the time derivative. Truncation error in the time differencing formula is of no concern since the solution is found when time derivatives are zero. The boundary conditions given by Eqs. (18) and (19) are imposed by using second order accurate backward differencing for χ_r at r = 1 and second order accurate forward differencing for χ_Θ at r = 1, Θ = 0. The resulting system of equations advances the value for χ from t to t + Δt. Eq. (20)establishes the initial values from which the relaxation procedure begins. Once the values are advanced from t to t + Δt throughout the entire interior mesh, and on the boundaries, E is evaluated from Eq. (21) and the iteration is ready to advance another time step. The time step is selected at each time level based on a heuristic approach to the stability question. By analogy with the two-dimensional diffusion equations in Cartesian coordinates studied by R. D. Richtmyer [2], the stability criterion for the current problem was selected as

$$(r^2\mathcal{B}^2 + M_\infty^2 A)\,\frac{\Delta t}{\Delta r^2} + (\mathcal{B}^2 + M_\infty^2 C)\,\frac{\Delta t}{\Delta\Theta^2} \leq \frac{1}{2}\ .$$

However, by running several numerical experiments it was discovered that stability could still be maintained if the 1/2 were increased to 0.61. Spatial mesh widths were defined by selecting 60 equally spaced rays with 10 equally spaced points on each ray, not including the point r = 0.

6. RESULTS

The above described numerical method has been programmed and run on a UNIVAC 1106 for the cases of symmetric flow past a Joukowsky profile, lifting flow past an 8.57% cambered Kármán-Trefftz airfoil of 10° trailing edge angle, and a NACA 0012 airfoil at 2° angle of attack. The latter solution is shown in Fig. 1. The iteration converged in all three cases for purely subsonic flow and sufficient accuracy was obtained after about 20 minutes of computer run time. For fully nonlinear, compressible flow past a Joukowsky or Kármán-Trefftz airfoil, other results for comparison with the current method could not be found in the literature. In Fig. 1, the results are compared with the numerical solution of Sells given by R. C. Lock [3]. The method can therefore be seen to work quite well and warrants further extension into the transonic range.

7. TRANSONIC FLOW CONSIDERATIONS

To accomplish this task, the substitution form given by Eq. (14), modified to account for the fully nonlinear potential equation, should be used to greatly reduce machine running time. The modification comes about because the coefficient of Φ_{yy} in Eq. (10a) is now a function of both dependent and independent variables whereas in the small disturbance theory, the coefficient is unity.

Thus the modification is derived by adding $f_1\Phi_{yt}$ and $(f_2/4)\Phi_{tt}$ to $f\Phi_{yy}$ where f, f_1, f_2 are functions of space and velocity, such that a parabolic equation results. Thus,

$$f_1^2 - 4f\frac{f_2}{4} = 0$$

or then

$$f_1^2 = f\, f_2$$

and, since f is known from (10a), one has an arbitrary choice to make for f_2 or f_1. Suppose one chooses $f_2 = f$, then the substitution term becomes

$$S(\Phi_t) = f\left(\frac{1}{4}\Phi_{tt} + \Phi_{yt}\right).$$

Another possibility would be to choose $f_2 = 1/f$, then the substitution term becomes

$$S(\Phi_t) = \frac{1}{f}\left(\frac{1}{4}\Phi_{tt} + f\Phi_{yt}\right).$$

The numerical difference scheme for spatial differences must change when the flow field becomes supersonic locally. Hence, an additional refinement in the method must be made so that the Mach number at each mesh point is determined and tested to see if the flow is supersonic or not. For supersonic points the space differences in the approximate streaming direction must be backward differences, not central differences as they are in the subsonic portion. For differences in the y-direction, the change over in difference scheme is not necessary. The difference scheme can also be changed to gain some artificial viscosity which, in addition to stability, furnishes an entropy inequality on the solution obtained. Shock waves should occur naturally as the solution progresses.

8. REFERENCES

[1] D. Catherall, D. N. Foster, C. C. L. Sells: Two-Dimensional Incompressible Flow Past a Lifting Aerofoil. RAE TR 69118 (1969).

[2] R. D. Richtmyer: Difference Methods for Initial-Value Problems. Interscience Publishers, Inc., New York, 1957, pp. 112-120.

[3] R. C. Lock: Test Cases for Numerical Methods in Two-Dimensional Transonic Flows. AGARD Rep. 575 (1970).

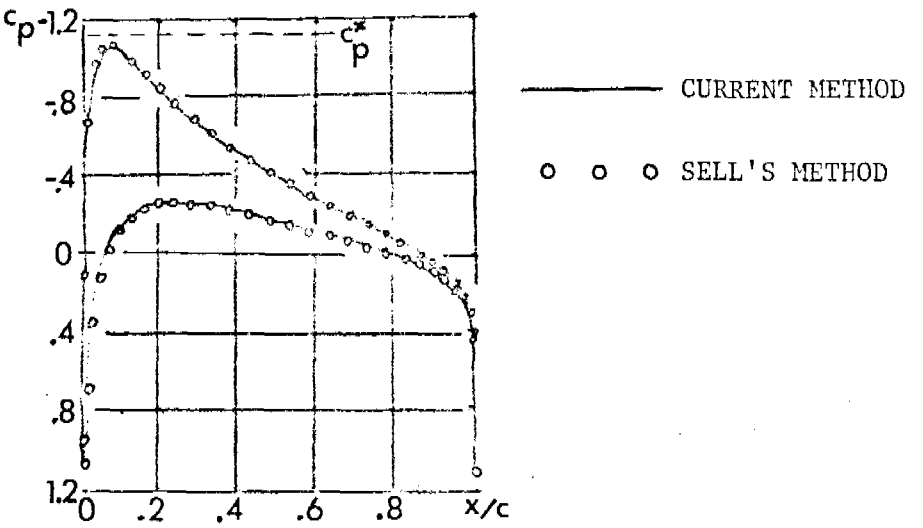


Fig 1. PRESSURE DISTRIBUTION ON AN NACA 0012 AIRFOIL AT $M_\infty = 0.63$, $\alpha = 2^o$

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

THE BACKWARD BEAM EQUATION AND THE NUMERICAL COMPUTATION OF DISSIPATIVE EQUATIONS BACKWARDS IN TIME

Alfred Carasso

Technical Summary Report #1534

March 1975

ABSTRACT

We present an expository survey of the backward beam equation approach in the numerical computation of parabolic equations backwards in time. We discuss linear and non-linear problems, and we present the details of several numerical experiments on problems where exact solutions are known. Our discussion includes problems with variable coefficients depending on time, as well as recently obtained results on the computation of the final value problem for Burgers' equation.

AMS (MOS) Subject Classifications - 35R25, 65M30

Key Words - Ill-posed problems; Numerical computation of backwards parabolic equations; The backward beam equation approach

Sponsored by the United States Army under Contract No. DA-31-124-ARO-D-462.

THE BACKWARD BEAM EQUATION AND THE NUMERICAL COMPUTATION OF DISSIPATIVE EQUATIONS BACKWARDS IN TIME † *

Alfred Carasso ‡

1. Introduction

The purpose of this paper is to survey the main results of [2-6], dealing with the development of a new algorithm for the approximate solution of backwards parabolic equations. The exposition is mostly without proofs, and the reader is referred to the original papers for a more detailed discussion. Since this work was begun, a substantial amount of computational experience has been accumulated. The method appears to be a powerful tool. Some computational experiments, undertaken by B. L. Buzbee at the Los Alamos Scientific Laboratories, have not been published. They are mentioned in Section 3. More recently, [6], the method was extended to an interesting example of a nonlinear equation, Burgers' equation, and extensive numerical experiments were carried out on problems for which exact solutions are known. In Section 5, we describe some of the results obtained in [6]. Further applications to nonlinear problems are currently under way.

† Sponsored by the National Science Foundation under NSF Grant GP 42536 and the United States Army under Contract No. DA-31-124-ARO-D-462.

‡ Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131.

* Expanded version of a talk given at the National Science Foundation Regional Conference on Ill-Posed Problems, University of New Mexico, May 1974.

2. Some general remarks on backwards dissipative equations

We shall focus attention on a special but quite interesting class of ill-posed problems, namely the problems which arise when the time direction is reversed in a dissipative evolution equation. Such equations distinguish a time direction as they describe irreversible phenomena. While the problem of determining the future from the present is well understood analytically, and a considerable literature exists which deals with the effective numerical computation of such forward problems, the same is not true of the backwards problem. Attempting to reconstruct the past from the present leads to tremendous difficulties, as the solutions depend discontinuously on the data, and this in an essential way. That is, continuous dependence cannot usually be restored by reconsidering the question in some other metrizable linear space. What seems necessary to restore continuity is the imposition of additional constraints on the class of admissible solutions, such as requiring the solutions to be positive or to satisfy an a priori bound. On the other hand, while the constrained problem is well posed in the analytical sense, there remains the question of devising effective numerical methods in which the constraints can be incorporated, if one wishes to approximate the solutions. Thus, the standard stable marching schemes (such as the Crank-Nicolson or other Padé approximants) which have been widely successful in the numerical computation of forward dissipative problems, are necessarily unstable when the time direction is reversed. This is a general theorem

to be found in [17, p. 59]. What is missing in these classical schemes is a means of incorporating the constraints. In recent years, it has become apparent that the essential difficulties in the above class of ill-posed problems are algorithmic in character. By and large, the numerical analysis of these problems is not as well developed as are the analytical questions such as backwards uniqueness and stability under a prescribed bound. Nevertheless, a great deal of effort and ingenuity has been applied towards the effective numerical computation of such problems; we refer to [16] for an extensive bibliography, and to the paper by Miller in the present volume for a viewpoint somewhat different from ours.

It should be pointed out that not all methods which have been proposed are equally effective or applicable. Thus, methods which require exact knowledge of the data are of limited usefulness in applications. Even when an analytical expression is available for the data, round-off error in digitizing that data plays a considerable role and its effect cannot be ignored. The same is true for schemes which require the data and corresponding solution to have a Fourier transform with compact support. As the spatial mesh is refined, extraneous high frequencies are injected into the solution by the rounding process. These high frequencies may be amplified without bound, in arbitrarily small time intervals, as time evolves backwards. Furthermore, if a smoothing process is used at each time step, it may not be possible to decide which parts of the spectrum should be filtered out. Thus, in nonlinear problems, or even in

linear problems with variable coefficients, there may be genuine interactions between different frequency bands. Several methods which have been proposed (for linear problems) begin by recasting the problem in the form of an integral equation, i.e.

$$S(T)u = v \tag{2.1}$$

where v is the given terminal data at time T, u is the desired initial data, and $S(T)$ is the solution operator at time T in the forward analytic problem. Since $[S(T)]^{-1}$ is unbounded, various regularization techniques are then employed to approximate u. The difficulty here is that except in very simple problems, one generally does not know $S(T)$ explicitly. Thus, if the equation has variable coefficients depending on time, one rarely has formulae for the fundamental solution. It is a happy fact that many stabilized backwards dissipative problems enjoy the property of Hölder-continuity with respect to the data on compact subintervals of $(0, T]$. Usually, the exponent $\mu(t)$ is a function of t which tends to zero as $t \downarrow 0$. As a simple example, in linear problems with a self adjoint operator one can give a sharp estimate for the error at time t, corresponding to an error δ at time T, given an a priori bound, M, for the initial data. One has in the L^2 norm,

$$\|\varepsilon(t)\| \leq 2\, M^{\frac{T-t}{T}}\, \delta^{\frac{t}{T}}, \quad 0 \leq t \leq T . \tag{2.2}$$

Here $\mu(t) = \frac{t}{T}$, and there is "destruction of information" as $t \downarrow 0$. This loss of information is more severe than that which takes place in the usual well-

posed linear problems of mathematical physics. On the other hand, the situation is far worse in other classes of stabilized ill-posed problems. Thus, in [9], F. John gives an example where a much weaker type of continuity, "logarithmic continuity", can be shown to actually hold. In that example, the data must be known to an accuracy of $10^{-400,000,000}$ in order to produce an accuracy of 10^{-3} in the corresponding solutions! An important requirement for a numerical method is that it should preserve the Hölder dependence inherent in the analytic problem. This requirement is stronger than stability. The latter simply requires that round-off or other errors in the data not be amplified without bound in finite time; however, stability alone may not prevent the scheme from amplifying errors in the data far beyond the theoretical limit set by the estimate (2.2), resulting in an unwarranted loss of precious information. Thus in [8], an interesting example is given of a stable marching scheme for the backwards heat equation, in which logarithmic continuity with respect to the data actually holds. Similarly, [18], Tichonov's method for the backward heat equation is only logarithmically continuous with respect to the data.

Even at the analytical level, there remain genuine difficulties in obtaining a priori stability estimates in backwards dissipative problems. Such estimates are of fundamental importance as they measure the rate at which information is destroyed as time evolves backwards. To illustrate this point, consider the situation for the Navier-Stokes equations backwards in time. In [12] the authors consider the class of smooth solutions

of the Navier Stokes equations, in a space time domain $\Omega \times [0, T]$ satisfying

$$(2.3) \qquad \sup_{(x,t)\in \Omega\times[0,T]} \{|u|^2 + |\mathrm{curl}\, u|^2 + |u_t|^2\} \leq N^2 ,$$

where N is a given a priori bound. Let ν be the kinematic viscosity, V the volume of Ω, and let $u_1(x,t)$, $u_2(x,t)$ be two smooth solutions of the Navier-Stokes equations, satisfying (2.3) in $\Omega \times [0, T]$, and such that,

$$(2.4) \qquad \|u_1(\cdot, T) - u_2(\cdot, T)\|^2_{L^2(\Omega)} \leq \delta .$$

It is shown in [12] that then, for $0 \leq t \leq T$,

$$(2.5) \qquad \|u_1(\cdot, t) - u_2(\cdot, t)\|^2_{L^2(\Omega)} \leq (4N^2)^{1-\mu(t)}\, \delta^{\mu(t)} \exp\left[\frac{N^4(t-\mu(t)T)}{\nu^2}\right]$$

where $\mu(t)$ is given by

$$(2.6) \qquad \mu(t) = \frac{1 - \exp\left[\frac{2(N^2+1)t}{\nu}\right]}{1 - \exp\left[\frac{2(N^2+1)T}{\nu}\right]} .$$

The estimate (2.5) establishes Hölder-continuous dependence on the data and implies backwards uniqueness of smooth solutions. On the other hand, as far as computing the solutions are concerned, (2.5) is rather disconcerting. For example, if $V = T = N = 1$, $\nu = 10^{-1}$, and $\delta = 10^{-50}$, we have from (2.5) at $t = T/2 = 1/2$,

$$(2.7) \qquad \|u_1(\cdot, 1/2) - u_2(\cdot, 1/2)\|^2_{L^2(\Omega)} \leq 4\, e^{50}(10^{-50})e^{-20} \approx 10^{22} .$$

Moreover, (2.7) is little changed by choosing $\delta = 10^{-500}$, since the rate at which $\mu(t) \downarrow 0$ is so large, even at such small Reynolds numbers. It is not known whether the exponent $\mu(t)$ is sharp, or whether constraints different from (2.3), possibly involving other combinations of derivatives, would result in a more encouraging estimate. In non-linear problems, a major task appears to be that of isolating classes of equations for which one can obtain fairly reasonable Hölder estimates. As a very small beginning, the one dimensional Burgers' equation,

$$u_t = \nu u_{xx} - uu_x + f(x,t), \qquad 0 \le x \le L, \qquad 0 \le t \le T \ , \tag{2.8}$$

is considered in [6]. It is straightforward to show that if $u_1(x,t)$, $u_2(x,t)$, are two solutions satisfying

$$\text{Max}\{|u_i|,\ |u_{i,t}|,\ |u_{i,tt}|,\ |u_{i,xt}|\} \le N, \qquad i = 1,\ 2 \ , \tag{2.9}$$

for $(x,t) \in [0,L] \times [0,T]$, and if

$$\|u_1(\cdot,T) - u_2(\cdot,T)\|_{L^2} \le \delta \ , \tag{2.10}$$

then for $0 \le t \le T$,

$$\|u_1(\cdot,t) - u_2(\cdot,t)\|_{L^2} \le 2K(t)\, N^{\frac{T-t}{T}}\, \delta^{\frac{t}{T}} \ , \tag{2.11}$$

where

$$K(t) = \exp\left[\frac{4NL + t(T-t)\{(NL)^2 + (1+3\nu)NL\}}{4\nu}\right]. \tag{2.12}$$

In [6], extensive numerical experiments are presented for the final value problem for Burgers' equation. Some of these results will be described in

the present paper later on. (See Section 5). One of the points suggested by these experiments is that the factor $K(t)$, which involves the Reynolds number, appears to play a significant role only when the solutions to Burgers' equation develop steep gradients approaching a "shock". For more reasonable solutions, one finds that, even in single precision, i.e. with a unit round-off error of about 10^{-8}, one can often attain significant accuracy at 90% of the way back from $T = 1$. Thus, the difficulty of reconstructing past steep gradients from future smoothed data appears to be explained by the factor $K(t)$ in (2.11). Conceivably, in the case of the Navier-Stokes equations, the similar exponential factor in (2.5) may alone suffice to account for the difficulty of reconstructing steep gradients, and in some suitable norm, there might well exist a Hölder estimate in which $\mu(t)$ is independent of the Reynolds number, and decays linearly with t.

3. The backward beam equation approach in self-adjoint problems with time independent coefficients

We shall now describe a new method, which was recently developed in [2], for computing linear self-adjoint parabolic equations backwards in time. In this section we consider the case where the spatial operator, A, is independent of t. The case where A depends on t is more subtle and is discussed in Section 4.

Let $f(x)$ be a given function in $L^2(\Omega)$, where Ω is a bounded domain in R^N, in R^N with a smooth boundary $\partial\Omega$. Let A be a non-negative self-adjoint operator in $L^2(\Omega)$; in the concrete cases, A is the unbounded operator corresponding to a self-adjoint elliptic boundary value problem in Ω, with, say, zero Dirichlet data on $\partial\Omega$. Given the positive constants δ, M, T, we consider the following problem.

Find all solutions of

$$u_t = -Au, \qquad 0 < t \leq T \ , \tag{3.1}$$

such that

$$\|u(\cdot, T) - f\| \leq \delta \ , \tag{3.2}$$

and

$$\|u(\cdot, 0)\| \leq M \ . \tag{3.3}$$

To solve this stabilized backwards problem, consider the following device. Set

$$k = \frac{1}{T} \log\left(\frac{M}{\delta}\right) \tag{3.4}$$

and then put $v = e^{kt}u$ in (3.1). This leads to

$$v_t = -(A-k)v, \qquad 0 < t \leq T \ , \tag{3.5}$$

$$\|v(\cdot, T) - e^{kT}f\| \leq e^{kT}\delta, \qquad \|v(0)\| \leq M \ . \tag{3.6}$$

Differentiating (3.5) with respect to t, we obtain the "backward beam equation" associated with (3.1), namely,

$$v_{tt} = Bv \ , \quad 0 < t \leq T \ , \tag{3.7}$$

where $B = (A-k)^2$. Thus, B is a non-negative self-adjoint operator in $L^2(\Omega)$; in particular, B is "m-accretive". For such equations, it is easy to show that solutions are norm-convex. We have,

$$\frac{d^2}{dt^2}\|v(t)\|^2 = 2\|v'(t)\|^2 + 2\,\mathrm{Re}(Bv, v) \geq 0 \ . \tag{3.8}$$

In particular, if v(t) is a solution of (3.7),

$$\|v(t)\| \leq \frac{T-t}{T}\|v(0)\| + \frac{t}{T}\|v(T)\| \ . \tag{3.9}$$

The last inequality suggests that the "initial-terminal value" or "two-point" problem is well-posed for (3.7), i.e. data should be prescribed at $t = 0$ and at $t = T$. Using Hadamard's classical example of the Cauchy problem for Laplace's equation, it is easily shown that the initial-value problem is in general ill-posed for (3.7). It also follows from the spectral representation of B, that solutions to the two point problem exist, for <u>arbitrary</u> data v(0), v(T) in $L^2(\Omega)$; moreover, as shown in [5], (3.7) has the "smoothing" property. That is, if A has

sufficiently smooth coefficients, arbitrarily high Sobolev norms of the solution at time t, $0 < t < T$, can be estimated in terms of the L^2 norms of the data $v(0)$, $v(T)$. In a very real sense, (3.7) is an "elliptic" equation in Hilbert space.

Let $w(\cdot, t)$ be the unique solution of

$$w_{tt} = Bw, \qquad 0 < t < T, \tag{3.10}$$

$$w(0) = 0 \qquad w(T) = e^{kT} f \ . \tag{3.11}$$

We then have

Theorem 3.1

Let $u(\cdot, t)$ be any solution of the stabilized backwards problem (3.1)-(3.3). Let k be as in (3.4), and let $w(\cdot, t)$ be the unique solution of (3.10)-(3.11). Then,

$$\| e^{-kt} w(\cdot, t) - u(\cdot, t) \| \leq M^{\frac{T-t}{T}} \delta^{\frac{t}{T}} \tag{3.12}$$

Moreover, if A has smooth coefficients, N is the dimension of Ω, q is a positive integer, and $2\sigma > \frac{N}{2} + q$, there is a constant K such that for $0 < t < T$,

$$\text{Max}_{|\beta| \leq q} \| D^\beta u(\cdot, t) - e^{-kt} D^\beta w(\cdot, t) \|_\infty \leq \tag{3.13}$$

$$K \left\{ (t)^{-\sigma} + (T-t)^{-\sigma} + \left(\frac{\log(\frac{M}{\delta})}{T} \right)^\sigma \right\} M^{\frac{T-t}{T}} \delta^{\frac{t}{T}} \ .$$

Proof:

If u is any solution of (3.1)-(3.3), $v = e^{kt}u$ satisfies (3.7) and the inequalities (3.6). Let $z(\cdot, t) = v - w$. Then, from (3.9) and (3.6), we have,

$$(3.14) \qquad \|e^{-kt} z(\cdot, t)\| \leq \frac{T-t}{T} M e^{-kt} + \frac{t}{T} e^{k(T-t)} \delta = M^{\frac{T-t}{T}} \delta^{\frac{t}{T}} .$$

This proves (3.12). The proof of (3.13) is more complicated and we refer the reader to [5].

Remark 1. Since the estimate (2.2) is sharp, for the difference of any two solutions of the backwards problem, it follows that $e^{-kt}w$ above, is a "best-possible" L^2 approximation to any solution of the backwards problem. Moreover, even though the data $f(x)$ is an approximation to $u(\cdot, T)$ in the L^2-norm, $e^{-kt}w$ approximates the solutions of the backwards problem, together with their derivatives, in the L^∞ norm, on $0 < t < T$.

Remark 2. By using (3.12) and the triangle inequality, we obtain an independent proof of the convexity estimate, (2.2), for the difference of any two solutions of the backwards problem. Similarly, (3.13) and the triangle inequality lead to a maximum norm stability estimate, for the derivatives of any two solutions, in terms of the L^2 norm of the data.

In actual numerical computation of the solution of (3.10)-(3.11), the time variable is discretized using a centered time discretization. With $T = (N+1)\Delta t$, we have,

$$(3.15) \qquad \frac{w^{n+1} - 2w^n + w^{n-1}}{\Delta t^2} = Bw^n, \qquad n = 1, 2, \ldots, N ,$$

$$(3.16) \qquad w^0 = 0 , \qquad w^{N+1} = e^{kT} f .$$

This system of linear equations can be written in tridiagonal matrix form, with an unbounded operator along the main diagonal. See [3], [4], [2]. Using the tridiagonal algorithm, one obtains the existence and uniqueness of solutions to (3.15), (3.16), together with the basic norm-convexity property (3.9), for the solution of this finite difference analog; the proof uses only the "m-accretiveness" of B. Moreover, this method of proof provides one of several possible algorithms for the solution of this system of linear equations. To discretize the spatial operator B, one uses finite difference analogs of the elliptic operator, or Galerkin methods using trial functions satisfying the boundary conditions. Either method preserves the "m-accretiveness" of B. Consequently, the fully-discrete scheme is unconditionally stable and has the property (3.9). Finally, estimates such as (3.12), (3.13), supplemented by the truncation error at time t in the fully-discrete scheme, remain valid for the solution of the fully discrete problem. We refer the reader to [2] for a more detailed discussion of these matters. By and large, the numerical analysis of (3.10), (3.11), is very similar to that for elliptic boundary value problems, for which an abundant literature exists. Thus, direct or iterative methods may be used to solve the system of linear equations. In [2], a computational example is discussed in detail for a one dimensional problem. Further examples will be given later in the present paper, when we discuss the backwards problem for Burgers' equation. (See Section 5). Finally, we mention some successful computations, on two-dimensional problems in rectangular regions, carried out by B. L. Buzbee at the Los Alamos Scientific Laboratories.

4. Self-adjoint problems with time dependent coefficients

An important feature of the backward beam equation approach is that the method is applicable to self-adjoint parabolic equations with smooth time dependent coefficients. The assumption of smooth dependence on time is important, as backwards uniqueness may fail for non-smooth coefficients. See the example of Miller in [14].

For each $t \geq 0$, let $a(t;u, v)$ be the symmetric bilinear form on $H_0^m(\Omega)$ given by

$$a(t;u, v) = \sum_{|p|,|q| \leq m} \int_\Omega a_{pq}(x, t)\, D^q u\, \overline{D^p v}\, dx \tag{4.1}$$

where the a_{pq} depend smoothly on x and t, and

$$a_{pq} = \overline{a}_{qp} \,. \tag{4.2}$$

We assume $a(t;u, v)$ to be strongly coercive on $H_0^m(\Omega)$, i.e. there exists a positive constant ω, independent of t, such that

$$a(t;v, v) \geq \omega \|v\|_m^2, \qquad \forall\, v \in H_0^m(\Omega) \,, \tag{4.3}$$

where $\| \quad \|_m$ denotes the Sobolev norm. Let $\dot{a}(t;u, v)$ be the bilinear form obtained from $a(t;u, v)$ by replacing a_{pq} by $\dot{a}_{pq} = \frac{\partial}{\partial t} a_{pq}$. The form $\dot{a}(t;u, v)$ will play an important role in the subsequent discussion. Let $P(t)$ be the unbounded self-adjoint operator in $L^2(\Omega)$ defined by $a(t;u, v)$, i.e.,

(4.4) $$\langle P(t)v, v\rangle = a(t; v, v) \quad \forall\, v \in H_0^m(\Omega)$$

where $\langle\ ,\ \rangle$ is the scalar product in $L^2(\Omega)$. An integration by parts shows that $P(t)$ corresponds to the self-adjoint elliptic partial differential operator,

(4.5) $$P_0(t) = \sum_{|p|,\,|q| \le m} (-1)^{|p|} D^p(a_{pq}(x,t) D^q u), \quad x \in \Omega,\ t > 0\ ,$$

together with the Dirichlet boundary conditions,

(4.6) $$D^p u = 0 \quad \text{on } \partial\Omega, \quad |p| \le m-1, \quad t \ge 0\ .$$

We shall consider the parabolic problem

(4.7) $$u_t = -P(t)u, \quad t > 0\ .$$

and we note that from (4.6), the domain of $P(t)$ is <u>fixed</u> as t varies. We now introduce the following definitions.

<u>Definition</u>

The parabolic problem (4.7) is <u>minimal-smoothing</u> on $[0, T]$ if

(4.8) $$\dot{a}(t; v, v) \le 0 \quad \forall\, v \in V, \quad 0 \le t \le T\ .$$

It is <u>strongly-smoothing</u> if

(4.9) $$\dot{a}(t; v, v) \le 2\gamma \|v\|^2, \quad \gamma > 0\ .$$

where $\|\ \|$ denotes the L^2 norm. We say that (4.7) is <u>maximal-smoothing</u> if

(4.10) $$\dot{a}(t; v, v) \leq \alpha \|v\|_m^2, \qquad \alpha > 0 \ .$$

The above definition distinguishes three broad classes of problems. Further refinements are clearly possible. In the interest of simplicity of exposition, these refinements are not considered here. It will be seen later that the maximal-smoothing case is the hardest to compute backwards in time. This difficulty is not a defect of our method. Rather, it is associated with the type of Hölder estimate which obtains in that case. Each of the maximal-smoothing and strongly-smoothing cases can be reduced to the minimal-smoothing case by means of a preliminary transformation.

In the maximal-smoothing case this reduction is accomplished by stretching the time variable. With ω and α the constants in (4.3) and (4.10), define the function

(4.11) $$\psi(s) = (\frac{\omega}{\alpha}) \log(1 + \alpha s/\omega), \qquad s \geq 0 \ .$$

Then, $\psi'(s) > 0$, $\psi''(s) < 0$ and

(4.12) $$\omega\psi'' + \alpha(\psi')^2 = 0 \ .$$

From the bilinear form $a(t; u, v)$ in (4.1), we construct the form $b(s; u, v)$ where

(4.13) $$b(s; u, v) = a(\psi(s); u, v)\, \psi'(s), \qquad s \geq 0 \ ,$$

$$= \sum_{|p|, |q| \leq m} \int_\Omega b_{pq}(x, s)\, D^q u\ \overline{D^p v}\, dx \ ,$$

with

$$b_{pq}(x, s) = a_{pq}(x, \psi(s))\, \psi'(s), \qquad s \geq 0 \quad . \tag{4.14}$$

Let $b'(s; u, v)$ be the symmetric bilinear form on $H_0^m(\Omega)$, obtained from (4.13) by replacing $b_{pq}(x, s)$ by $\dfrac{\partial b_{pq}}{\partial s}$. We then have,

$$b'(s; u, v) = \psi''\, a(\psi(s); u, v) + (\psi')^2\, \dot{a}(\psi(s); u, v) \quad . \tag{4.15}$$

Hence, using (4.3), (4.10), (4.12) and the fact that $\psi'' < 0$, we obtain from (4.15),

$$b'(s; v, v) \leq [\alpha(\psi')^2 + \omega\psi'']\, \|v\|_m^2 = 0, \quad \forall\, v \in H_0^m(\Omega) \quad . \tag{4.16}$$

We now put $t = \psi(s)$ in the parabolic problem (4.7). Let

$$\xi(x, s) = u(x, \psi(s)), \quad x \in \Omega, \quad s \geq 0 \ ; \tag{4.17}$$

then, ξ satisfies the parabolic problem,

$$\xi_s = -G_0(s)\xi, \quad x \in \Omega, \quad s > 0 \ , \tag{4.18}$$

$$D^p \xi = 0, \quad x \in \partial\Omega, \quad s \geq 0, \quad |p| \leq m - 1 \ , \tag{4.19}$$

where $G_0(s) = \psi'(s)\, P_0(\psi(s))$. This transformed parabolic problem is the one generated by the symmetric bilinear form $b(s; u, v)$. Since $b'(s; v, v) \leq 0$, we have the minimal-smoothing case for the transformed problem.

In the <u>strongly smoothing</u> case, we put

$$b(t; u, v) = a(t; u, v) - 2\gamma t \int_\Omega u \bar{v}\, dx \quad . \tag{4.20}$$

Then, from (4.9), $\dot{b}(t; v, v) \leq 0$. Next, put

$$\xi(x, t) = e^{\gamma t^2} u(x, t), \quad x \in \Omega, \quad t > 0 . \tag{4.21}$$

in the parabolic problem (4.7). Then ξ satisfies,

$$\xi_t = -G_0(t)\xi, \quad x \in \Omega, \quad t > 0 , \tag{4.22}$$

$$D^p \xi = 0, \quad x \in \partial\Omega, \quad t \geq 0, \quad |p| \leq m - 1 , \tag{4.23}$$

where $G_0(t) = P_0(t) - 2\gamma t$. This is the problem generated by the bilinear form $b(t; u, v)$ in (4.20). Thus, (4.21), transforms the strongly smoothing case into the minimally smoothing case.

For the purposes of the following discussion, it may now be assumed without loss of generality that the parabolic problem,

$$u_t = -P(t)\, u, \quad t > 0 , \tag{4.24}$$

is minimally smoothing on the interval $[0, T]$. Given $f(x)$ in $L^2(\Omega)$, and the positive constants $\tilde{\delta}$, M, T, consider the following problem. Find all solutions of (4.24) such that

$$\|u(\cdot, T) - f\| \leq \tilde{\delta} \tag{4.25}$$

$$\|u(\cdot, 0)\| \leq M . \tag{4.26}$$

As in Section 3, we put $\tilde{k} = \frac{1}{T} \log\left(\frac{M}{\tilde{\delta}}\right)$ and $v = e^{\tilde{k}t} u$ in (4.24), to obtain,

$$v_t = -(P(t) - \tilde{k})v, \quad 0 < t < T . \tag{4.27}$$

Differentiating with respect to t, we obtain

(4.28) $$v_{tt} = A(t)v, \qquad 0 < t < T,$$

where $A(t)$ is an unbounded self-adjoint operator in $L^2(\Omega)$. See [4, Section 3]. In fact, for each fixed $t > 0$, $A(t)$ is the unbounded operator corresponding to the following elliptic boundary value problem of order $4m$ in Ω:

(4.29) $$[(P_0(t) - \tilde{k})^2 - \dot{P}_0(t)]u = 0, \quad x \in \Omega,$$

(4.30) $$D^p u = D^p[P_0(t)u] = 0, \quad x \in \partial\Omega, \quad |p| \leq m - 1;$$

where $\dot{P}_0(t)$ is the differential operator obtained from $P_0(t)$ in (4.5) by differentiating the coefficients with respect to t. Since the problem (4.24) is minimally-smoothing by hypothesis, we have,

(4.31) $$\langle \dot{P}_0(t)v, v\rangle \leq 0, \quad \forall v \in H_0^m(\Omega).$$

Consequently, $A(t)$ in (4.28) is a non-negative self-adjoint operator for each t. Note, however, that even though (4.24) is a fixed domain parabolic problem, the backward beam equation (4.28) now involves a variable domain operator, $A(t)$, in general. See [4, Section 3]. Unlike the problem in Section 3, it is no longer possible to prove existence theorems for (4.28) by using the spectral representation of $A(t)$ at each t. Remarkably enough, one can obtain strong results, (i.e. existence of solutions lying in the domain of $A(t)$ for each t), even in this variable domain case, for the two-point problem associated with the finite difference analog of (4.28). No assumptions need be made about the manner in which $D_A(t)$ varies with

t, and the proof uses only the "m-accretiveness" of $A(t)$ for each t. Let $T = (N+1)\Delta t$, and consider the system of difference equations,

$$\frac{w^{n+1} - 2w^n + w^{n-1}}{\Delta t^2} - A^n w^n = 0, \qquad n = 1, 2, \ldots, N \ , \tag{4.32}$$

$$w^0 = a, \qquad w^{N+1} = b \ , \tag{4.33}$$

where $a, b \in L^2(\Omega)$, and $A^n \equiv A(n\Delta t)$. In [3], the following theorem is proved.

Theorem 4.1

There exists a unique solution, $w(t)$, in (4.32), (4.33), with $w(t) \in D_A(t)$, for each $t = n\Delta t$, $n = 1, 2, \ldots, N$, and this for arbitrary $a, b \in L^2(\Omega)$. Moreover,

$$\|w(t)\| \leq \frac{T-t}{T} \|a\| + \frac{t}{T} \|b\| \ . \tag{4.34}$$

The proof of Theorem 4.1 given in [3] is constructive, and is based on the tridiagonal algorithm. If finite element methods are used to discretize $A(t)$ for each t, such methods preserve the accretiveness of $A(t)$. Consequently, the resulting fully-discrete scheme also has the norm convexity property (4.34). In practice, iterative methods, such as block relaxation techniques, may be used to solve the block tridiagonal system of linear equations. See [4].

We now turn to the question of approximating the solutions to the parabolic problem (4.7), backwards in time, given an a priori bound, M,

for the initial data, and given a function $f(x) \in L^2(\Omega)$, such that $\|u(\cdot, T) - f\| \leq \delta$. If (4.7) is minimally-smoothing, we solve the system (4.32) with the two-point conditions,

$$w^0 = 0, \quad w^{N+1} = e^{kT} f \ , \tag{4.35}$$

and

$$k = \frac{1}{T} \log \frac{M}{\delta} \ . \tag{4.36}$$

We then define,

$$u_{app}(t) = e^{-kt} w(t) \tag{4.37}$$

as our approximation to the solutions of the backwards problem. Thus, this case is treated in exactly the same way as the problem in Section 3.

If (4.7) is strongly-smoothing, we first transform to the minimal case by means of (4.21), and solve the resulting system (4.32) for the transformed problem, with the two-point conditions,

$$w^0 = 0, \quad w^{N+1} = e^{\gamma T^2} e^{kT} f \ , \tag{4.38}$$

where,

$$k = \frac{1}{T} \{\log M - \log(e^{\gamma T^2} \delta)\} \ . \tag{4.39}$$

We then define,

$$u_{app}(t) = e^{-\gamma t^2} e^{-kt} w(t) \tag{4.40}$$

as our approximation to the solutions of the original problem. Finally, in the maximal-smoothing case, we transform to the stretched variable, s, as in (4.17). Let $\Delta s = S/N+1$, where

$$S = \left(\frac{\omega}{\alpha}\right)\left[e^{\alpha T/\omega} - 1\right] . \tag{4.41}$$

We solve the system (4.32) in the s-variable with the two-point conditions,

$$w^0 = 0, \qquad w^{N+1} = e^{kS_f} \tag{4.42}$$

and $k = \frac{1}{S}\log\left(\frac{M}{\delta}\right)$.

Let

$$\psi^{-1}(t) = \frac{\omega}{\alpha}\left[e^{\alpha t/\omega} - 1\right] . \tag{4.43}$$

As our approximation to the solutions of the original problem, we define,

$$u_{app}(t) = \exp[-k\psi^{-1}(t)]\; w(\psi^{-1}(t)) , \tag{4.44}$$

where $w(s)$ is the solution of (4.32), (4.42). We then have the following result.

Theorem 4.2

Let $u(t)$ be any solution of the stabilized backwards problem for (4.7), and consider $u_{app}(t)$. In the minimally smoothing case, we have,

$$\|u(t) - u_{app}(t)\| \leq M^{\frac{T-t}{T}}\,\delta^{\frac{t}{T}} + O(\Delta t^2) . \tag{4.45}$$

In the strongly-smoothing case,

$$\|u(t) - u_{app}(t)\| \leq \exp[\gamma t(T-t)]\, M^{\frac{T-t}{T}}\, \delta^{\frac{t}{T}} + O(\Delta t^2) \quad . \tag{4.46}$$

Finally, in the maximal smoothing case,

$$\|u(t) - u_{app}(t)\| \leq M^{1-\mu(t)}\, \delta^{\mu(t)} + O(\Delta t^2) \quad , \tag{4.47}$$

where,

$$\mu(t) = \frac{e^{\alpha t/\omega} - 1}{e^{\alpha T/\omega} - 1} \quad . \tag{4.48}$$

Proof

The proof of each of the three inequalities follows from the norm-convex property of the solution of the backward beam equation associated with each of the transformed problems, as in the proof of (3.12) in Theorem 3.1, together with a subsequent inverse transformation in the case of the last two inequalities. See [2]. The extra term $O(\Delta t^2)$ represents the combined spatial and time discretization errors, as we are now considering the fully discrete scheme as opposed to the evolution equation (4.28). It is assumed that the spatial mesh is chosen so that the spatial truncation error is of the same magnitude as that of the time discretization.

Remark. By making $\Delta t \to 0$ in the above error estimates, we recover logarithmic convexity estimates originally obtained by Agmon-Nirenberg, [1]. Note the exponential decay to zero as $t \downarrow 0$, of the exponent $\mu(t)$ in (4.47). An example of the maximal-smoothing case is provided by a

simple diffusion equation in which the diffusion coefficient grows with time. The strongly smoothing case corresponds to a diffusion equation with a growing zero order term. The minimal case corresponds to a constant or decaying diffusion coefficient. By considering a diffusion coefficient dependending only on t, one can show that (4.47) is sharp. In such a problem, considerably more precision in measurement is necessary at time T, in order to attain significant accuracy backwards in time, as compared with the other two cases.

5. Computing small solutions of Burgers' equation backwards in time

We shall now describe some recent results dealing with the application of the backward beam method in the computation of the final value problem for Burgers' equation. The reader is referred to [6], for proofs and a more detailed discussion of the main results. We consider the following initial boundary value problem for the one dimensional Burgers' equation,

$$u_t = \nu u_{xx} - uu_x + f(x,t), \quad 0 \le x \le L, \quad 0 \le t \le T \ , \tag{5.1}$$

$$u(0,t) = u(L,t) = 0, \quad t \ge 0 \ , \tag{5.2}$$

$$u(x,0) = a(x), \quad 0 \le x \le L \ , \tag{5.3}$$

where $a(x)$ and $f(x,t)$ are sufficiently smooth that the (unique) solution of (5.1)-(5.3) has sufficiently many derivatives on $[0,T]$. Let A be the positive self-adjoint operator corresponding to $-u''$ with zero boundary conditions, and let

$$F(u) = -uu_x \ . \tag{5.4}$$

We may then write (5.1)-(5.3) in the form of an evolution equation in $L^2[0,L]$, viz,

$$u_t = -\nu Au + F(u) + f(t), \qquad 0 < t < T \ , \tag{5.5}$$

$$u(0) = a \ . \tag{5.6}$$

One way of solving the forward problem is by means of the following iterative procedure,

$$u_t^0 = -\nu A u^0 + f(t), \qquad 0 < t < T \ , \tag{5.7}$$

$$u^0(0) = a \ , \tag{5.8}$$

and for each $m = 1, 2, 3, \ldots,$

$$u_t^m = -\nu A u^m + F(u^{m-1}) + f(t), \qquad 0 < t < T \ , \tag{5.9}$$

$$u^m(0) = a \ . \tag{5.10}$$

In fact, this procedure was used by Kato and Fujita in [10], [11], as a means of proving existence and uniqueness theorems for the Navier-Stokes equations, which they viewed as an initial value problem in Hilbert space. An important feature of the above iteration is that one proceeds through a sequence of inhomogeneous linear parabolic problems with constant coefficients. Concerning the convergence of this iterative process, we have the following theorem. See [6].

Theorem 5.1

Let $a(x)$ belong to $D(A^{\frac{1}{2}})$, and let $f(t) \in D(A^{\frac{1}{2}})$ with $\|A^{\frac{1}{2}} f(t)\| \in L^1[0, T]$. Let

$$\left(\frac{64\ LT}{\nu}\right)^{\frac{1}{2}} \left[\|A^{\frac{1}{2}} a\| + \int_0^T \|A^{\frac{1}{2}} f(s)\| ds\right] < 1 \ ; \tag{5.11}$$

then $A^{\frac{1}{2}} u^m$ exists on $[0, T]$ for every m. Let $\theta = 1 - \left[1 - \left(\frac{64\ LT}{\nu}\right)^{\frac{1}{2}} \|\|A^{\frac{1}{2}} u^0\|\|\right]^{\frac{1}{2}}$. Then $0 < \theta < 1$, and

(5.12) $$|||A^{\frac{1}{2}}u^m||| \leq \theta\left(\frac{\nu}{16LT}\right)^{\frac{1}{2}} .$$

Moreover, if $u(t)$ is the unique solution of (5.5), (5.6),

(5.13) $$|||A^{\frac{1}{2}}(u^m-u)||| \leq \theta^{m+2}\left(\frac{\nu}{64LT}\right)^{\frac{1}{2}} .$$

In the above statement of the theorem, $|||v|||$ denotes $\underset{0\leq t\leq T}{\text{Sup}}\,\|v(t)\|$, and $\|A^{\frac{1}{2}}v(t)\|^2 = \int_0^L |v_x(x,t)|^2\,dx$.

The condition (5.11) is sufficient for convergence in the norm $|||\quad|||$. In practice, convergence may occur even if (5.11) is severely violated, by as much as a factor of 1000 in some cases. See [6] for several numerical examples. On the other hand, (5.11) cannot be too severely violated for convergence to occur on $[0,T]$. In general, the convergence of (5.7)-(5.10) is only local in time, even if a unique smooth solution exists for all $t > 0$. An example of divergence, except for t sufficiently small, is given in [6]. In that example, the exact solution is known and is found to develop steep gradients, i.e. a smooth approximation to a "shock" evolves from the initial data. Convergence occurs before the gradients become too steep. See also Example 2 and Figure 1 below.

The essential idea behind the algorithm for the backwards problem, is to solve each linear parabolic problem in the Kato-Fujita sequence of iterates, backwards in time, using the value of the solution to (5.5), (5.6) at time T . Each such linear backwards problem is solved via the backward beam equation discussed in Section 3. Since each linear problem has

constant coefficients, the "Fourier Method" of Kreiss-Oliger can be used to great advantage. That is, one calculates the spatial derivatives at the mesh points, by differentiating the trigonometric polynomial which interpolates the function values at equally spaced grid points, using at least 2 points per significant wave length. If the data, inhomogeneous term, and solution have smooth periodic extensions, this technique for differentiation is highly accurate as shown in [13]. It is also extremely attractive in that one can make use of Fast Fourier Transform algorithms. (Such a method has been employed by Orszag in [15], for the Navier-Stokes equations.) Finally, with this particular technique for discretizing the space variable, the algebraic problem of inverting the "block tri-diagonal" matrix in (3.15) becomes trivial. One has a scalar positive definite tridiagonal matrix to invert for each Fourier component in turn, and such matrices can be efficiently inverted by a standard algorithm. It should be remarked that the high accuracy in spatial discretization is quite important in the present equation, where there may be high frequency components in the solution at positive times, even if such components are absent in the initial data. This is a basic property of the homogeneous equation. See the classic paper by Julian Cole in [7].

To motivate our main result, consider the following "thought experiment". Let $\{u^m(t)\}_{m=0}^{\infty}$ be the sequence of iterates in the forward problem, and let $u^m(T) = b^m$. Imagine that each $b_m(x)$ is known approximately. Let $\tilde{b}_m$ be the approximate value of b_m, and suppose

(5.14) $$\|A^{\frac{1}{2}}(b_m - \tilde{b}_m)\| \leq \delta, \qquad m = 0, 1, 2, \ldots ;$$

Suppose further that (5.11) is satisfied for the forward iteration. Then, from (5.12), we have

(5.15) $$|||A^{\frac{1}{2}}u^m||| \leq M = \theta\left(\frac{\nu}{64\,LT}\right)^{\frac{1}{2}} .$$

We may then pose the following linear backwards parabolic problem for each iterate $u^m(t)$:

Find all solutions of

(5.16) $$u_t^m = -\nu A u^m + F(u^{m-1}) + f(t), \qquad 0 < t \leq T ,$$

such that

(5.17) $$\|A^{\frac{1}{2}}(\tilde{b}_m - u^m(T))\| \leq \delta$$

(5.18) $$\|A^{\frac{1}{2}}u^m(0)\| \leq M .$$

Using the backward beam equation with $k = \frac{1}{T}\log(\frac{M}{\delta})$, each $u^m(t)$ may then be approximately solved backwards in time, and used to generate a new approximate inhomogeneous term for the next iteration. Because of the "destruction of information" as $t \downarrow 0$, it is clear that sizable errors will be generated at $t = 0$ and passed on to the next successive iterate. The next theorem assesses the accumulated error after m steps of this process. See [6] for the proof.

Theorem 5.2

Let (5.11) be satisfied. Let $w^m(t)$ be the sequence of successive approximations obtained via the backward beam equation. Let $u(t)$ be the solution of (5.5), (5.6). Then there exists a positive constant C_m, depending only on m, such that

$$(5.19)\quad \|A^{\frac{1}{2}}w^m(t) - A^{\frac{1}{2}}u(t)\| \leq C_m[\log(\tfrac{M}{\delta})]^{\beta_m} M^{\frac{T-t}{T}} \delta^{\frac{t}{T}} + \theta^{m+2}(\tfrac{\nu}{64\,LT})^{\frac{1}{2}}$$

where $\beta_m = 2^{m-3} + \frac{1}{2}$.

Remark. It follows from (5.19) that given any $\varepsilon > 0$, one can make $\|A^{\frac{1}{2}}w^m(t) - A^{\frac{1}{2}}u(t)\| < \varepsilon$, uniformly on compact subintervals of $(0, T]$, by choosing δ sufficiently small and m sufficiently large. On the other hand, even though (5.11) is satisfied so that the forward iterates, $u^m(t)$, converge on $[0,T]$, the above inequality does not imply convergence, for fixed $\delta > 0$, as $m \to \infty$.

The above theorem is not strictly applicable to the algorithm which is used in practice. In the first place, (5.11) may be severely violated. In the second place, the functions $\tilde{b}_m(x)$ are not available. Rather, one has an approximation $\tilde{b}(x)$ to the terminal value $u(x, T)$, where u is the solution of the non-linear problem (5.1)-(5.3), and,

$$(5.20)\qquad \|A^{\frac{1}{2}}u(\cdot, T) - A^{\frac{1}{2}}\tilde{b}\| \leq \delta \ .$$

From physical or other considerations, one may obtain an a priori bound,

$$(5.21)\qquad |||A^{\frac{1}{2}}u||| \leq M \ .$$

Using the given data $\tilde{b}(x)$ and setting $k = \frac{1}{T}\log(\frac{M}{\delta})$, each inhomogeneous linear parabolic problem is then solved backwards with the backward beam equation. In an extensive series of numerical experiments, with problems for which exact solutions are known, it is then found that Theorem 5.2 is qualitatively correct as far as the algorithm which is used in practice is

concerned. Thus, the solutions are well approximated after a relatively small number of iterations. Further iterations may lead to rapid divergence, even though the forward iteration converges. In many cases, one observes considerable improvement before the onset of divergence. The situation is analogous to that of the divergent power series in the theory of asymptotic expansions, where the first few terms often provide excellent approximations. Furthermore, in several experiements, the distance back into the past where significant accuracy can be attained, is greater than might be expected from the a priori stability estimate for Burgers' equation given in Section 2. We shall now give several numerical examples.

Example 1

Consider

$$u_t = \nu u_{xx} - uu_x, \qquad 0 < x < \pi, \qquad 0 < t < 1 \ , \tag{5.22}$$

$$u(0, t) = u(\pi, t) = 0, \qquad 0 \leq t \leq 1 \ , \tag{5.23}$$

$$u(x, 0) = u_0 \operatorname{Sin} x, \qquad 0 \leq x \leq \pi \ . \tag{5.24}$$

For any positive ν and any u_0, the exact solution of this problem was obtained by Cole in [7]. It is given by

$$u(x, t) = \frac{4\nu \sum_{n=1}^{\infty} e^{-\nu n^2 t} \, n \, I_n(\frac{u_0}{2\nu}) \operatorname{Sin} n x}{I_0(\frac{u_0}{2\nu}) + 2 \sum_{n=1}^{\infty} e^{-\nu n^2 t} \, I_n(\frac{u_0}{2\nu}) \operatorname{Cos} n x} \tag{5.25}$$

where $I_n(z)$ is the modified Bessel function of the first kind. From (5.25),

we observe the effect of non-linearity, in that the initial pure sine wave evolves into a periodic function in which all frequencies are present. It is instructive to associate a Reynolds number with the above problem. Following Cole, [7], we define

$$\text{Re} = \frac{u_0 \pi}{\nu} \quad . \tag{5.26}$$

In this example we chose $u_0 = 1$ and $\nu = .0025$, so that $\text{Re} = 126$. As far as the sufficient condition for convergence of the forward iteration is concerned, we actually have, in the present case,

$$\left(\frac{64\,LT}{\nu}\right)^{\frac{1}{2}} \left\| A^{\frac{1}{2}} a \right\| \approx 36 \quad . \tag{5.27}$$

The expression (5.25) was evaluated at 64 equally spaced mesh points on $[0, 2\pi]$, at $T = 1$, to generate the terminal data. The backward beam method was then used with $\Delta t = \frac{1}{301}$, and with Fourier techniques to discretize the space variable, using 64 equally spaced points on the period interval $[0, 2\pi]$. The computations were performed in single precision on UNIVAC equipment at the University of Wisconsin. Thus, the unit round-off error is of the order of 10^{-8}. Using $M = .1$, we then obtain $k = \log(\frac{M}{\delta}) \approx 18.7$.

Since the exact solution is known, a comparison of the computed solution with the exact solution was made after each successive iteration. At each time $t = n\Delta t$, $n = 1, 2, \ldots, N$, the relative error in the discrete spatial L^2 norm, was computed after each iteration, to observe the

behavior of the backwards iteration. This error is tabulated in Table 1, for the first six iterations, for 17 values of t lying between zero and 1. Little change in the relative error pattern appears after the fourth iteration. Despite the moderately large Reynolds number of 126 in the present example, and the appearance of the exponential factor exp[Re] in the Hölder estimate for Burgers' equation, we see that even with δ of the order of $.1 \times 10^{-8} = 10^{-9}$, a relative error of less than 10% is achieved as far as 93% of the way back from $T = 1$. In the above example, the exact solution does not develop steep gradients within the time interval $[0,1]$. Moreover, the forward iteration converges, although (5.11) is violated.

Example 2

The problem is again (5.22)-(5.24) but with $u_0 = 40$ and $\nu = 1$. As in the previous example,

(5.28) $$\mathrm{Re} = 126 .$$

However, in lieu of (5.11), we now have

(5.29) $$\left(\frac{64\,LT}{\nu}\right)^{\frac{1}{2}} \|A^{\frac{1}{2}}a\| \approx 711 .$$

In this example, the forward iteration diverges for $t \geq .04$. An independent evaluation of the exact solution, using (5.25), reveals that the initial sine wave develops steep gradients almost immediately. See Figure 1. A smooth approximation to a "shock" evolves, broadens, and then dies out. Attempts were made to compute this problem backwards in time, in

TABLE 1.

Relative error in the L^2 norm, as a function of time and number of iterations, in the backwards computation of Example 1

No of items / TIME	1	2	3	4	5	6
.0332	.29+00	.29+00	.29+00	.29+00	.29+00	.29+00
.0664	.95-01	.89-01	.85-01	.84-01	.84-01	.84-01
.0997	.54-01	.43-01	.35-01	.30-01	.30-01	.30-01
.1661	.46-01	.34-01	.26-01	.21-01	.21-01	.22-01
.2326	.42-01	.28-01	.22-01	.19-01	.19-01	.19-01
.2990	.37-01	.22-01	.17-01	.15-01	.16-01	.16-01
.3654	.34-01	.21-01	.17-01	.16-01	.17-01	.17-01
.4319	.29-01	.16-01	.14-01	.13-01	.13-01	.13-01
.4983	.27-01	.11-01	.84-02	.80-02	.81-02	.81-02
.5648	.23-01	.93-02	.79-02	.77-02	.78-02	.78-02
.6312	.18-01	.75-02	.66-02	.65-02	.65-02	.65-02
.6977	.17-01	.13-01	.12-01	.12-01	.12-01	.12-01
.7641	.12-01	.29-02	.26-02	.26-02	.26-02	.26-02
.8306	.82-02	.40-02	.39-02	.39-02	.39-02	.39-02
.8970	.83-02	.75-02	.75-02	.75-02	.75-02	.75-02
.9634	.45-02	.47-02	.47-02	.47-02	.47-02	.47-02
.9967	.32-02	.32-02	.32-02	.32-02	.32-02	.32-02

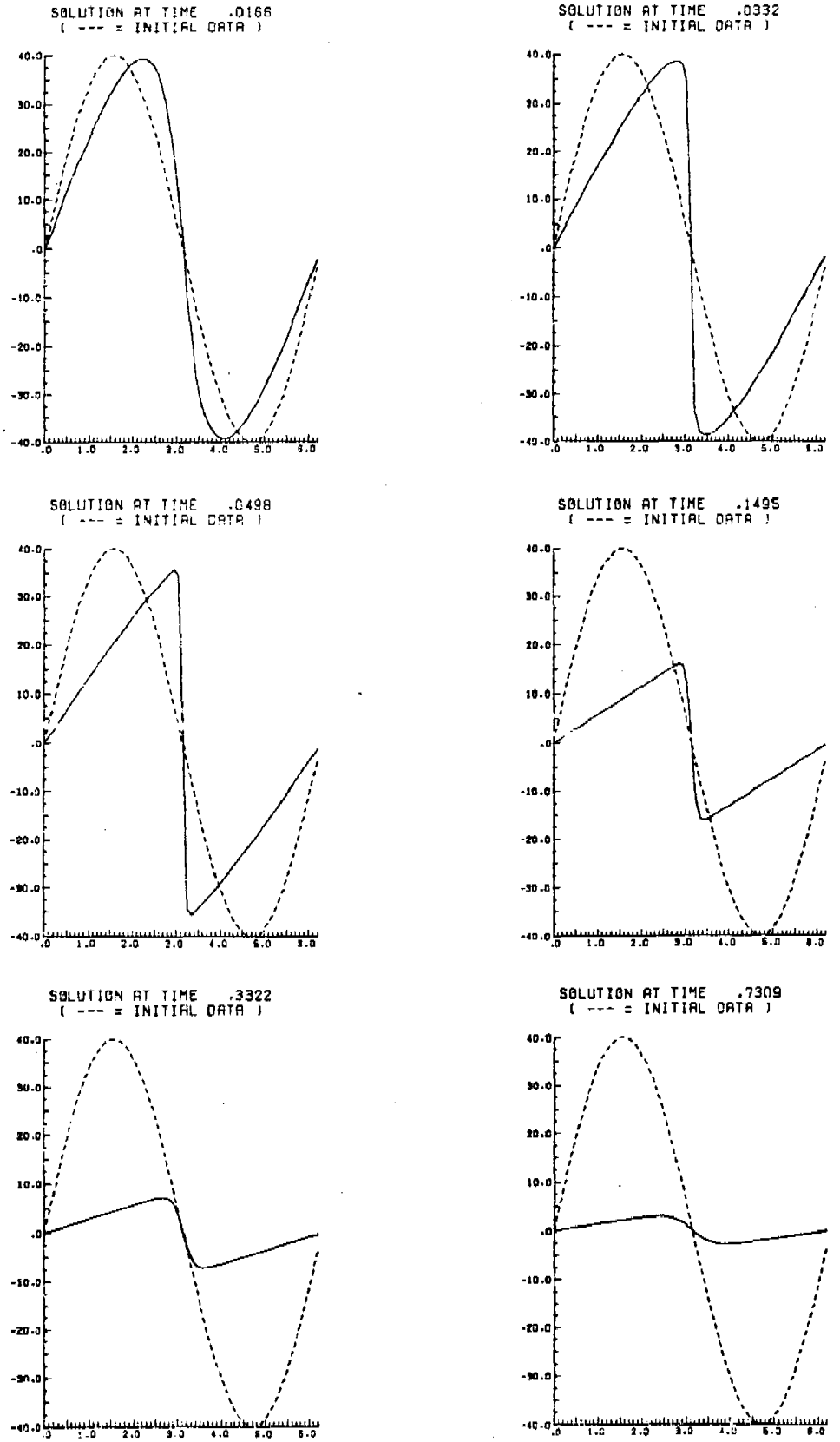


Figure 1. Development of steep gradients in the exact solution of the problem in Example 2.

single precision, starting from a time T_1 sufficiently close to zero, that the "shock" would still be clearly defined in the terminal data. This amounts to reconstructing steep gradients after they have been smoothed, and clearly requires considerable precision in measurement. No measurable success was achieved in this experiement.

Although Example 1 has the same Reynolds number as the present example, it appears that the exponential factor in the stability estimate plays a much more important role in the present problem.

Example 3

We now consider an inhomogeneous problem,

$$u_t = \nu u_{xx} - uu_x + 9\pi e^{-8\pi^2 \nu t} \operatorname{Sin} 4\pi x, \qquad 0 < x < 1, \quad 0 < t < 1, \tag{5.30}$$

$$u(0, t) = u(1, t) = 0, \quad t \geq 0, \tag{5.31}$$

$$u(x, 0) = 3 \operatorname{Sin} 2\pi x, \qquad 0 \leq x \leq 1. \tag{5.32}$$

The exact solution is

$$u(x.t) = 3e^{-4\pi^2 \nu t} \operatorname{Sin} 2\pi x . \tag{5.33}$$

We chose $\nu = 3/14$ so that $Re = 14$ in this experiement. Note however that in lieu of (5.11), we have

$$\left(\frac{64\,LT}{\nu}\right)^{\frac{1}{2}} \left[\| A^{\frac{1}{2}} a \| + \int_0^T \| A^{\frac{1}{2}} f(s) \| ds\right] \approx 487 . \tag{5.34}$$

Convergence of the forward iteration occurs at this and even higher values of (5.34) in this type of example. However, even here, where the solution does not develop steep gradients, divergence of the forward iteration occurs when $\nu = .05$, in which case (5.34) has a value of 2700. See [6]. The importance of the value of (5.34), rather than the Reynolds number, is quite apparent in the behavior of either the forward or backwards iterations, and this remains valid in all our experiments.

In the backwards computation of this problem, the relative error at 93% of the way back from $T = 1$, was found to be of the order of 500% after the first iteration! This error was then reduced to less than 10% after six iterations. At 77% of the way back from $T = 1$, the initial relative error of 110% was reduced to less than .3% after six iterations. In Figure 2, the first seven iterates are plotted, together with the exact solution, at $t = .0664$, the 93% value. In Figure 3, the iterates at 77% of the way back are depicted. Again the influence of the exponential factor in the convexity estimate, does not seem to be present in this single precision computation. Indeed, the accuracy which one can achieve at such long distances into the past is very encouraging.

4. An example of "asymptotic" convergence

In Example 3, the value of 487 in (5.34) is somewhat of a critical value insofar as observing the divergence phenomenon suggested by Theorem 5.2, in less than eight or nine iterations. We consider now the problem in Example 3 with a slightly lower value of ν, $\nu = 3/14.256789$.

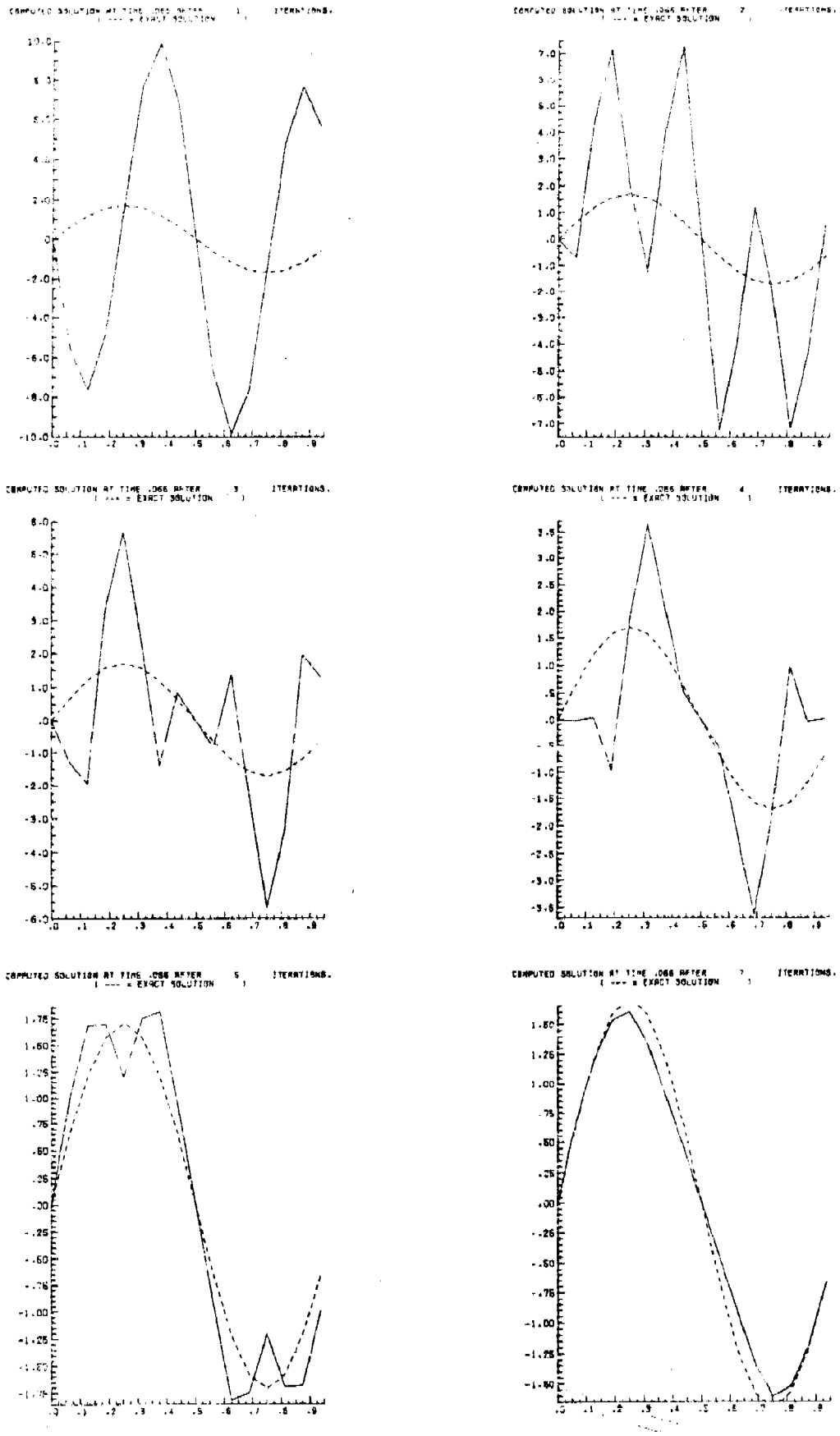


Figure 2. Convergence of the iteration at 93% of the way back from T = 1, in the computation of Example 3 backwards in time.

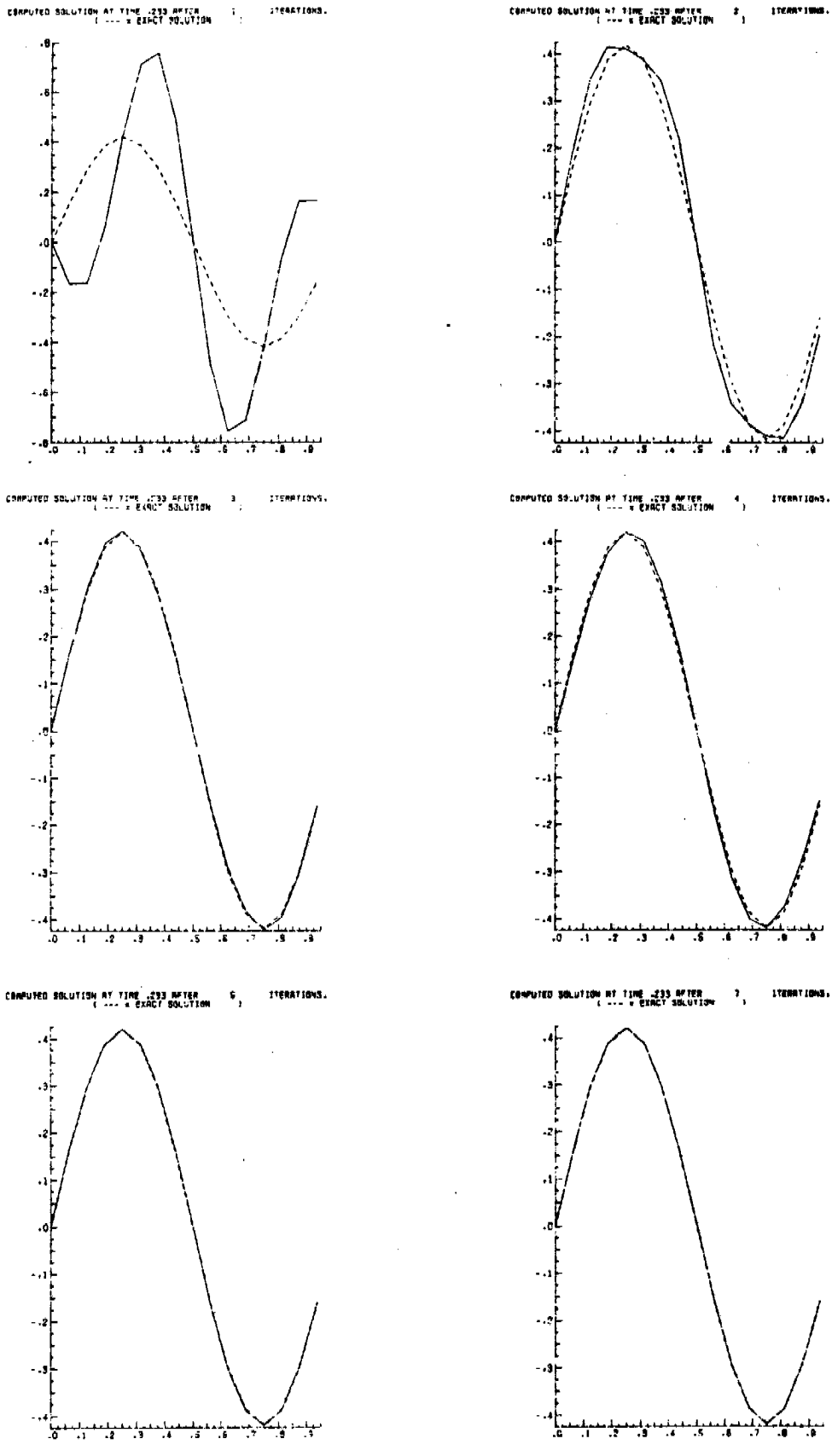


Figure 3. Convergence of the iteration at 77% of the way back from T = 1, in the computation of Example 3 backwards in time.

We then have,

$$(5.35) \qquad \left(\frac{64\,LT}{\nu}\right)^{\frac{1}{2}} \left[\|A^{\frac{1}{2}}a\| + \int_0^T \|A^{\frac{1}{2}}f(s)\|ds\right] = 496 \ .$$

In Figure 4, the first seven iterates at 77% of the way back from $T = 1$ are depicted. The third iteration gives the closest agreement. A more detailed discussion of this and other features of the algorithm is given in [6]. The forward iteration converges in this example.

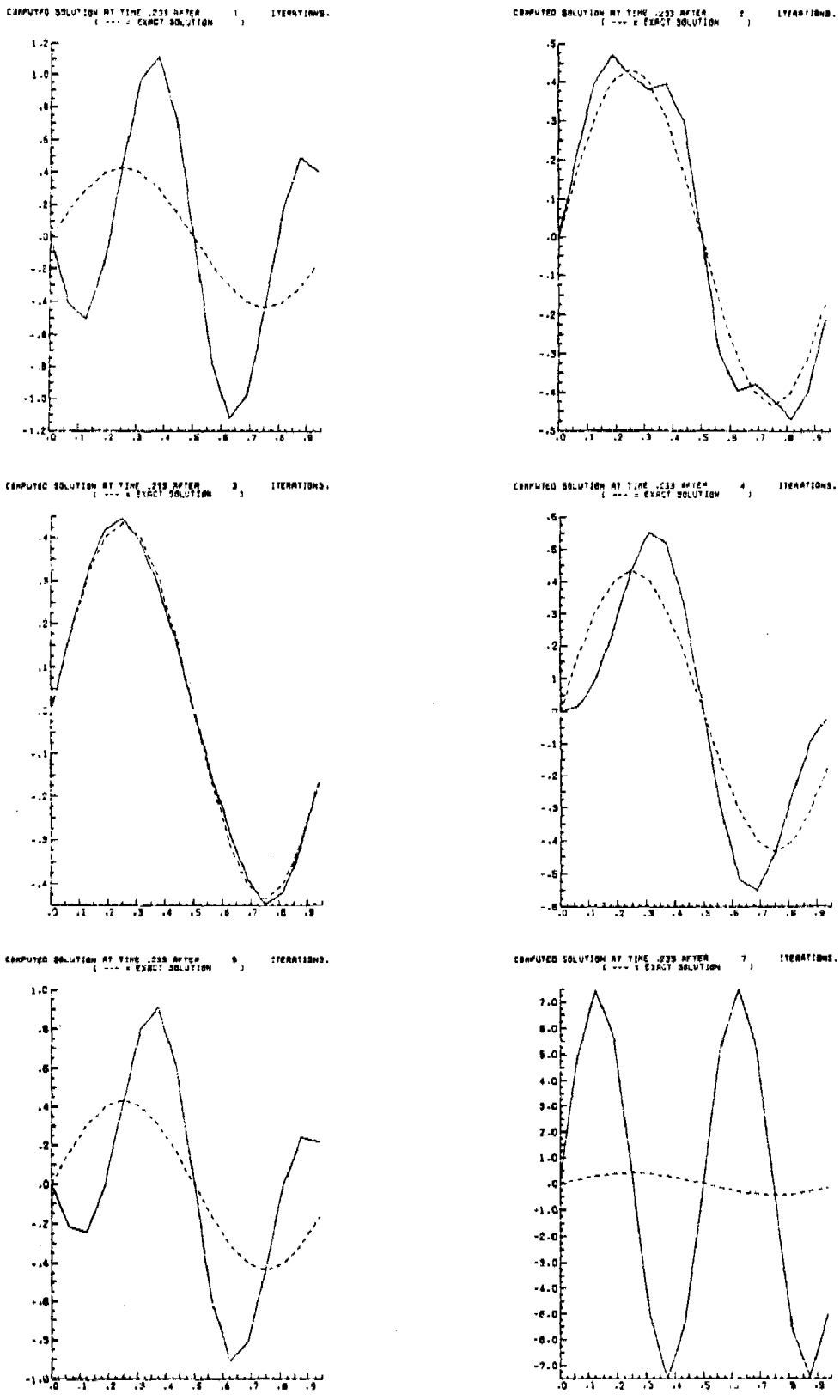


Figure 4. Asymptotic convergence phenomenon at 77% of the way back from T = 1, in the computation of Example 4 backwards in time.

1. S. Agmon and L. Nirenberg, Properties of solution of ordinary differential equations in Banach space, C. P. A. M. 16 (1963), 121-139.

2. B. L. Buzbee and A. Carasso, On the numerical computation of parabolic problems for preceding times, Math. Comp. 27 (1973), 237-266.

3. A. Carasso, The abstract backward beam equation, SIAM J. Math. Anal. 2 (1971), 193-212.

4. A. Carasso, The backward beam equation: Two A-stable schemes for parabolic problems, SIAM J. Numer. Anal. 9 (1972), 406-434.

5. A. Carasso, Error bounds in the final value problem for the heat equation, MRC Technical Summary Report #1479, September 1974, Mathematics Research Center, University of Wisconsin-Madison, Wis. 53706.

6. A. Carasso, Computing small solutions of Burgers' equation backwards in time, MRC Technical Summary Report #1525, January 1975, Mathematics Research Center, University of Wisconsin-Madison, Madison, Wis. 53706.

7. J. D. Cole, On a quasilinear parabolic equation occurring in aerodynamics, Quart. Appl. Math. 9 (1951), 225-236.

8. R. E. Ewing, The approximation of certain parabolic equations backwards in time by Sobolev equations, SIAM J. Math. Anal. (To appear).

9. F. John, Numerical solution of problems which are not well-posed in the sense of Hadamard, Proc. Rome Symp. Prov. Int. Comp. Center (1959), 103-116.

10. T. Kato and H. Fujita, On the non-stationary Navier-Stokes system, Rend. Sem. Mat. Univ. Padova 32 (1962), 243-260.

11. T. Kato and H. Fujita, On the Navier-Stokes initial-value problem, Arch. Rat. Mech. Anal. 16 (1965), 269-315.

12. R. J. Knops and L. E. Payne, On the stability of solutions of the Navier-Stokes equations backwards in time, Arch. Rat. Mech. Anal. 29 (1968), 331-335.

13. H. O. Kreiss and J. Oliger, Comparison of accurate methods for the integration of hyperbolic equations, Technical Report, No. 36, Dept. of Computer Sciences, Uppsala University, October 1971, Uppsala, Sweden.

14. K. Miller, Nonunique continuation for uniformly parabolic and elliptic equations in self-adjoint divergence form with Hölder continuous coefficients, Arch. Rat. Mech. Anal. 54 (1974), 105-117.

15. S. A. Orszag, Numerical simulation of incompressible flows within simple boundaries I, Studies in Applied Mathematics 50 (1971), 293-327.

16. L. E. Payne, Improperly posed problems in partial differential equations, Lecture Notes, National Science Foundation Regional Conference on Ill-Posed Problems, May 1974; Department of Mathematics and Statistics, University of New Mexico, Albuquerque, N. M. 87131.

17. R. D. Richtmyer and K. W. Morton, Difference Methods for Initial Value Problems, 2nd Ed., Interscience, New York, (1967).

18. J. N. Franklin, On Tichonov's Method for Ill-Posed Problems, Math. Comp. 28 (1974), 889-907.

SPECIAL SOLUTIONS OF THE ONE-DIMENSIONAL PARABOLIC EQUATION

Siegfried H. Lehnigk
Physical Sciences Directorate
US Army Missile Research, Development and Engineering Laboratory
US Army Missile Command
Redstone Arsenal, AL 35809

ABSTRACT. From a mathematical point of view, two classes of conservative diffusion processes are discussed which can be described by means of a similarity variable which depends linearly on the original space variable.

1. INTRODUCTION. We are interested in finding solutions $Q(x,t)$ of the parabolic equation

$$A(x)\, Q_{xx} + B(x)\, Q_x + C(x)\, Q = Q_t \tag{1.1}$$

in the domain D: $x > 0$, $t > 0$, with the property

$$Q(x,t) \in L^1[0,\infty) \quad \forall t > 0$$

and, in particular,

$$0 \leq \int_0^\infty Q(x,t)\, dx = \text{const} < \infty \ .$$

Such solutions are called conservative.

The problem of conservative solutions has a long history. We give three typical references: Fourier [1], Doetsch [2], and Feller [3].

We assume that $A,B,C \in C(0,\infty)$. Equation (1.1) is a Fokker-Planck equation if and only if $C = B' - A''$. In this case, it can be written in the form

$$[AQ]_{xx} - [\tilde{B}Q]_x = Q_t$$

with $\tilde{B} = 2A' - B$.

Let us consider two examples:

Ex.1. (1.1) with $A = \alpha x$, $B = \beta_1 + \beta_2 x$, $C = \beta_2$, $\alpha > 0$, $\beta_{1,2} \in R$ (Fokker-Planck). This equation has been thoroughly investigated by Feller [3]. In D it has the particular solution

3. TWO SETS OF COEFFICIENT FUNCTIONS. We now give two sets of coefficient functions A(x), B(x), and C(x) for equation (2.1) which satisfy the conditions (2.6) and (2.8). Proofs and all details will be presented elsewhere.

$$(3.7)\qquad Q(x,t) = b^{-1}(t)\,\xi^{-1}\left(\exp -\frac{1}{2}\alpha^{-1}\beta_1 \log^2 \xi\right) \times \left[C_1\, {}_1F_1\left(\frac{1}{2}(2-\sigma), \frac{3}{2}; \frac{1}{2}\alpha^{-1}\beta_1 \log^2 \xi\right) \log \xi + C_2\, {}_1F_1\left(\frac{1}{2}(1-\sigma), \frac{1}{2}; \frac{1}{2}\alpha^{-1}\beta_1 \log^2 \xi\right)\right]$$

$$C_{1,2} = \text{const}\qquad,\quad \xi = xb^{-1}(t)$$

where b(t) is determined by (3.6) under the initial conditions $t_o = 0$, $b_o > 0$, with $\varkappa = 3\alpha - \beta_2$, and

$$\int_0^\infty Q(x,t)\,dx = \begin{cases} 0 \text{ if } \sigma > 0 \\ C_2\sqrt{2\pi\alpha\beta_1^{-1}} \text{ if } \sigma = 1 \text{ (Fokker-Planck)} \end{cases} .$$

4. INITIAL AND BOUNDARY BEHAVIOR. It is appropriate to conclude with a brief remark on the initial and boundary behavior of the solutions covered by Theorems A and B.

In general diffusion processes in the domain $x > 0$, $t > 0$, it does not make sense to approach the origin (0,0) because of possible discontinuities in the initial and boundary behavior of solutions at that point. Therefore, one has to consider the general initial-boundary value problem [2] with perpendicular approach to the boundaries of the domain $x > 0$, $t > 0$, which excludes the approach to (0,0). From this point of view, a solution Q(x,t) of (2.1) is called singular if

$$Q(x,t) \to 0 \text{ as } t \downarrow 0 \text{ for fixed } x \in (0,\infty)$$

and

$$Q(x,t) \to 0 \text{ as } x \downarrow 0 \text{ for fixed } t \in (0,\infty) .$$

This terminology is due to Doetsch [5] (see also [2]).

The conservative solutions (3.3) of (2.1) with coefficients (3.1) are singular if $b_o = 0$ and if either

1) $\lambda < 1$, $1 + \lambda - \alpha^{-1}\beta_1 > 0$,

or

2) $\lambda > 1$.

The existence of singular solutions is rather disturbing within the framework of diffusion theory and its applications since they introduce nonuniqueness of the general initial-boundary value problem.

The solutions (3.7) of equation (2.1) with coefficients (3.5) are not singular since b_0 is positive. They do not go to 0 as $t \downarrow 0$.

REFERENCES

1. J. B. J. Fourier, Théorie analytique dela Chaleur, OEuvres de Fourier, vol. 1, Gauthier-Villars, Paris (1888).

2. G. Doetsch, Handbuch der Laplace-Transformation, vol. 3, Birkhäuser, Basel (1956).

3. W. Feller, Ann. Math. 54, 173 (1951).

4. S. Kepinski, Math. Ann. 61, 397 (1905).

5. G. Doetsch, Math. Z. 22, 293 (1925).

INTEGRATION OF $\int_0^\infty F(x)J_0(ax)J_1(bx)dx$

Shunsuke Takagi
U.S. Army Cold Regions Research And
Engineering Laboratory, Hanover, NH

INTRODUCTION

Infinite integrals involving Bessel functions under the integral sign are of extreme importance in many branches of mathematical physics. We encountered several types of the integrals of the form

$$\int_0^\infty F(x)\ J_0(ax)\ J_1(bx)\ dx \qquad (1)$$

in the study of the viscoelastic deformation of an ice plate floating on water (Ref. 1).

When $F(x)$ is an even function and has only algebraic singularities (i.e. poles and branch points), the integral (1) can be transformed, as shown later, to a contour integral. Therefore, when poles only constitute the singularities, the application of the residue theorem enables us to integrate it.

There are many other more complicated integrals that cannot be simply integrated by use of the residue theorem. A simple example is a variant of (1) where $F(x)$ has a branch point. A more complicated example is the one where $F(x)$ has an essential singularity; the application of the contour integral is out of the question in this case. We may hope, however, that all these difficulties can be resolved by use of Barnes' integral representations.

Watson states that "the Bessel function under the integral sign may be replaced by the contour integral of Barnes' type involving Gamma functions, and the order of the integration is then changed; this very powerful method has not previously been investigated in a systematic

manner" (Ref. 2, p. 383). Watson (Ref. 2), however, does not develop this method to cover all the needs arising from practical application. He rather imposes severe restrictions, as shown later, to avoid ambiguous applications.

It is shown in this paper that his restrictions may be removed. However, as shown later, different forms of Barnes' representations do not necessarily yield one and the same result. Watson's restrictions extricate us from this trouble, although they prohibit the application of this method to a majority of the interesting cases.

The objective of this paper is to present the difficulties we have encountered, and to solicit theoretical mathematicians to solve them. It is my hope that the mysteries enshrouding the application of the Barnes' representations may be lifted in the near future and unreserved use of them will be guaranteed.

CONTOUR INTEGRAL

The following theorem gives the conditions that enable us to transform (1) to a contour integral.

Theorem: Let $F(z)$ be an even function of z that has only algebraic singularities (i.e. poles and branch points) in the upper half plane of the complex variable $z = x + iy$. Then,

$$\int_0^\infty F(x)\, J_1(ax)\, J_0(bx)\, dx \tag{1}$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} F(z)\, H_1^{(1)}(az)\, J_0(bz)\, dz \quad , \text{ when } a>b \tag{2}$$

$$= \frac{1}{2}\oint_{-\infty}^{\infty} F(z)\, J_1(az)\, H_0^{(1)}(bz)\, dz \quad , \text{ when } a<b \qquad (3)$$

where $\oint_{-\infty}^{\infty}$ means the integral passing through the contour shown in Figure 1. The value at $a=b$ is given by the average of the limits approached from the region $a>b$ and the region $a<b$.

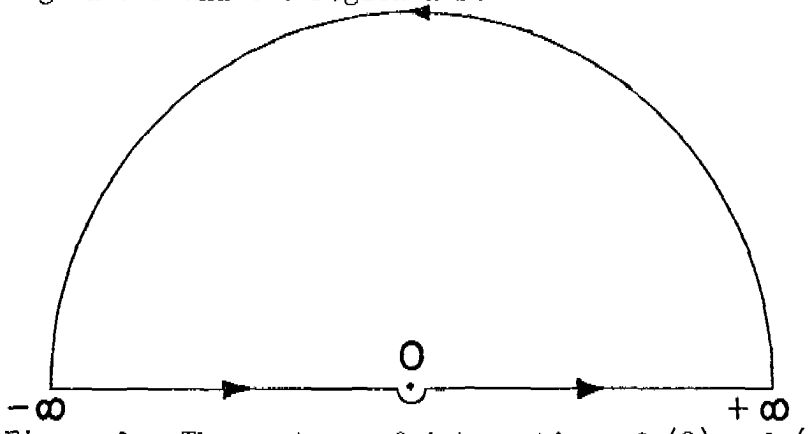


Figure 1. The contour of integration of (2) and (3).

Proof: Use of the relations

$$H_1^{(1)}(-z) = H_1^{(2)}(z)$$

and

$$J_0(-z) = J_0(z)$$

shows that

$$\frac{1}{2}\int_{-\infty}^{\infty} F(z)\, H_1^{(1)}(az)\, J_1(bz)\, dz = \frac{1}{2}\int_0^{\infty} F(z)\left[H_1^{(1)}(az) + H_1^{(2)}(az)\right] J_0(bz)\, dz \qquad (a)$$

The right hand side of (a) is equal to (1).

Use of the asymptotic formulas

$$H_1^{(1)}(az) \sim \sqrt{\frac{2}{\pi az}}\, e^{i[az - (3\pi/4)]}$$

and

$$J_0(bz) \sim \sqrt{\frac{2}{\pi bz}}\, \cos\left(bz - \frac{\pi}{4}\right)$$

shows that the left hand side of (a) is equal to (2). The case $a>b$ is thus proved. The case $a<b$ can be similarly proved.

The value at $a=b$ may be computed by use of Barnes' theory as done by Watson (see Ref. 2, p. 402). But this case is not essential to our analysis; we may not try to go through the complete analysis of this case in the present paper.

Example 1.

$$\int_0^\infty \frac{1}{1+x^4} J_1(ax)\, J_0(bx)\, dx \tag{4}$$

$$= \ker' a \operatorname{ber} b - \operatorname{kei}' a \operatorname{bei} b + \frac{1}{a} \quad \text{when } a \geq b \tag{5}$$

$$= \operatorname{ber}' a \ker b - \operatorname{bei}' a \operatorname{kei} b \quad \text{when } a \leq b \tag{6}$$

Proof: When $a>b$, the integral (4) transforms to $I(a>b)$ defined below,

$$I(a>b) = \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{1+z^4} H_1^{(1)}(az)\, J_0(bz)\, dz$$

The poles in the upper half plane are $e^{\pi i/4}$, $e^{3\pi i/4}$, and zero. Therefore

$$I(a>b) = \pi i \operatorname{Res}(e^{\pi i/4}) + \pi i \operatorname{Res}(e^{3\pi i/4}) + \frac{\pi i}{2} \operatorname{Res}(0)$$

Note that the pole at $z = 0$ is enclosed with a semicircle, as shown in Figure 1. Thus we find:

$$I(a>b) = \pi i \frac{1}{4} e^{\frac{-3\pi i}{4}} H_1^{(1)}(ae^{\frac{\pi i}{4}})\, J_0(be^{\frac{\pi i}{4}}) +$$

$$+ \pi i \frac{1}{4} e^{\frac{-9\pi i}{4}} H_1^{(1)}(ae^{\frac{3\pi i}{4}})\, J_0(be^{\frac{3\pi i}{4}}) +$$

$$+ \frac{\pi i}{2} \cdot \left(-\frac{2i}{\pi a}\right)$$

Substituting the relations,

$$H_1^{(1)}(xe^{\frac{3\pi i}{4}}) = \frac{2}{\pi i}(\mathrm{ker}_1 x + i\,\mathrm{kei}_1 x)$$

$$J_o(xe^{\frac{3\pi i}{4}}) = \mathrm{ber}_o x + i\,\mathrm{bei}_o x$$

$$H_1^{(1)}(xe^{\frac{\pi i}{4}}) = -\frac{2}{\pi i}(\mathrm{ker}_1 x - \mathrm{kei}_1 x)$$

and

$$J_o(xe^{\frac{\pi i}{4}}) = \mathrm{ber}_o x - i\,\mathrm{bei}_o x$$

we find

$$I(a>b) = \frac{1}{a} + \frac{1}{\sqrt{2}}\left\{(\mathrm{ker}_1 a + \mathrm{kei}_1 a)\ \mathrm{ber}_o b + (\mathrm{ker}_1 a - \mathrm{kei}_1 a)\ \mathrm{bei}_o b)\right\}$$

Using the relations

$$\mathrm{ker}_1 x + \mathrm{kei}_1 x = \sqrt{2}\ \mathrm{ker}'_o x$$

and

$$-\mathrm{ker}_1 x + \mathrm{kei}_1 x = \sqrt{2}\ \mathrm{kei}'_o x$$

$I(a>b)$ reduces to (5), where we have dropped the suffix zero.

When $a<b$, integral (4) transforms to $I(a<b)$ defined below,

$$I(a<b) = \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{1+z^4} J_1(az)\ H_o^{(1)}(bz)\ dz$$

The poles in the upper half plane are $e^{\pi i/4}$ and $e^{3\pi i/4}$. Therefore

$$I(a<b) = \pi i\ \mathrm{Res}(e^{\pi i/4}) + \pi i \mathrm{Res}(e^{3\pi i/4})$$

Thus we find:

$$I(a<b) = \pi i\ \frac{1}{4}\ e^{\frac{-3\pi i}{4}} J_1(ae^{\frac{\pi i}{4}})\ H_o^{(1)}(be^{\frac{\pi i}{4}})$$

$$+ \pi i\ \frac{1}{4}\ e^{\frac{-9\pi i}{4}} J_1(ae^{\frac{3\pi i}{4}})\ H_o^{(1)}(be^{\frac{3\pi i}{4}})$$

Substituting the relations,

$$H_0^{(1)}(xe^{\frac{3\pi i}{4}}) = \frac{2}{\pi i}(\mathrm{ker}_0 x + i\,\mathrm{kei}_0 x)$$

$$J_1(xe^{\frac{3\pi i}{4}}) = \mathrm{ber}_1 x + i\,\mathrm{bei}_1 x$$

$$H_0^{(1)}(xe^{\frac{\pi i}{4}}) = \frac{2}{\pi i}(\mathrm{ker}_0 x - i\,\mathrm{kei}_0 x)$$

$$J_1(xe^{\frac{\pi i}{4}}) = -(\mathrm{ber}_1 x - i\,\mathrm{ber}_1 x)$$

we find

$$I(a<b) = \frac{1}{\sqrt{2}}\left\{(\mathrm{ber}_1 a + \mathrm{bei}_1 a)\mathrm{ker}b + (\mathrm{ber}_1 a - \mathrm{bei}_1 b)\,\mathrm{kei}b\right\}$$

Using the relations

$$\mathrm{ber}_1 x + \mathrm{bei}_1 x = \sqrt{2}\,\mathrm{ber}'x$$

$$-\mathrm{ber}_1 x + \mathrm{bei}_1 x = \sqrt{2}\,\mathrm{bei}'x$$

$I(a<b)$ reduces to (6), where we have dropped the suffix zero.

Formulas (5) and (6) are continuous at $a=b$. To prove this, note that

$$w_1(x) = \mathrm{ber}x + i\,\mathrm{bei}x$$

and

$$w_2(x) = \mathrm{ker}x + i\,\mathrm{kei}x$$

are the solution of

$$x^2\frac{d^2w}{dx^2} + x\frac{dw}{dx} - ix^2w = 0$$

The real part of the Wronskian

$$\begin{vmatrix} w_1 & w_2 \\ w_1' & w_2' \end{vmatrix} = -\frac{1}{x}$$

gives the continuity of (5) and (6) at $a=b$.

BARNES' INTEGRAL REPRESENTATIONS

Proposition 1.

Barnes' representations of $J_\nu(x)$,

$$J_\nu(x) = \frac{1}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\Gamma(-s)\ (\frac{1}{2}x)^{\nu+2s}}{\Gamma(\nu+s+1)}\, ds \tag{7}$$

is valid for $\nu \geq 0$ and $x \geq 0$. When $\nu=0$, the order of the integration and the substituting $x=0$ cannot be exchanged.

Watson's restriction (Ref. 2, p. 192) that $R(\nu)>0$ and $x>0$ may be loosened, although ν is restricted in this proposition to real numbers in order to state simply the strange property at $x=0$. Note that Watson prohibits the use of (7) for $J_o(x)$. This restriction is removed here.

Proof.

We consider

$$L_1 = \lim_{|s|\to\infty} \frac{\Gamma(-s)\ (\frac{x}{2})^{\nu+2s}}{\Gamma(\nu+s+1)}\ s \tag{a}$$

where $0 \leq x < \infty$, in order to show that the contour integral derived from (7) by drawing the semicircle of infinitely large radius to the right of the imaginary axis amounts to zero.

Changing $-s$ in $\Gamma(-s)$ to $+s$ by using the formula

$$\Gamma(-z) = \frac{-\pi}{\Gamma(z+1)\sin\pi z} \tag{8}$$

where z is a complex number that is neither zero nor positive, L_1 becomes

$$L_1 = \lim_{|s|\to\infty} \frac{-\pi}{\Gamma(s)\sin\pi s} \frac{(x/2)^{\nu+2s}}{\Gamma(\nu+s+1)} \tag{b}$$

Letting

$$s = re^{i\theta} \tag{9}$$

we get

$$\lim_{r\to\infty} |\sin\pi s| = \lim_{r\to\infty} \frac{1}{2} Max(e^{i\pi s}, e^{-i\pi s}) = \lim_{r\to\infty} \frac{1}{2} e^{\pi r|\sin\theta|} \tag{c}$$

Substituting (c) and the asymptotic expansions of $\Gamma(s)$ and $\Gamma(\nu+s+1)$, (b) becomes

$$L_1 = \lim_{r\to\infty} |-e^{A_1}| \tag{d}$$

where

$$A_1 = (\nu+2s)\log\frac{x}{2} + (\nu+2s+1) - \pi r|\sin\theta| - (s-\tfrac{1}{2})\log s - (\nu+s+\tfrac{1}{2})\log(\nu+s+1)$$

which we transform to

$$A_1 = (\nu+2s)\log\frac{ex}{2s} - (\nu+s+\tfrac{1}{2})\log\frac{\nu+s+1}{s} - \pi r|\sin\theta| + 1 \tag{e}$$

Taking the real part of A_1, we find

$$L_1 = \lim_{r\to\infty} \left(\frac{ex}{2r}\right)^{\nu+2r\cos\theta} \exp[r(2\theta\sin\theta - \pi|\sin\theta|) + 1]$$

Because

$$2\theta\sin\theta - \pi|\sin\theta| \leq 0$$

and

$$\nu + 2r\cos\theta \geq 0$$

in the range

$$-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$$

it is proved that

$$L_1 = \text{finite} \quad \text{when } \nu = 0 \text{ and } \theta = \frac{\pi}{2}$$
$$= 0 \qquad \text{in the other cases.}$$

Therefore we can conclude that Barnes' theory by use of (7) is valid for $\nu \geq 0$, unless, when $\nu=0$, the order of the integration and the substituting $x=0$ is exchanged.

Proposition 2.

Barnes' representation of $H_\nu^{(1)}(x)$,

$$\pi e^{\frac{1}{2}(\nu+1)\pi i} H_\nu^{(1)}(z) = \frac{1}{2\pi i}\int_{-R(\nu)-\infty i}^{-R(\nu)+\infty i} \Gamma(-\nu-s)\Gamma(-s)\left(-\frac{1}{2}iz\right)^{\nu+2s} ds \qquad (10)$$

is valid for $|\arg(-iz)| \leq \frac{\pi}{2}$ and $\nu \geq 0$. The order of the integration and the substituting $z=0$ cannot be exchanged.

Watson's restriction (Ref. 2, p. 192) that $|\arg(-z)| < \frac{\pi}{2}$ is made exact in the above, although ν is restricted in this proposition to real numbers in order to state simply the strange property at $z=0$. Note that Watson prohibits the use of (10) for real values of z. This restriction is removed here.

Proof. we consider

$$L_2 = \lim_{|s|\to\infty} \left|\Gamma(-\nu-s)\Gamma(-s)\left(-\frac{iz}{2}\right)^{\nu+2s} s\right| \qquad (a)$$

in order to show that the contour integral derived from (10) by drawing the semicircle of infinitely large radius to the right of the imaginary axis amounts to zero.

Changing the negative arguments of $\Gamma(-\nu - s)$ and $\Gamma(-s)$ to the positive arguments by use of (8), L_2 becomes

$$L_2 = \lim_{|s|\to\infty} \left| \frac{\pi^2}{\Gamma(1+\nu+s)\Gamma(s)} \frac{\left(\frac{-iz}{2}\right)^{\nu+2s}}{\sin\pi(\nu+s)\sin\pi s} \right| \qquad (b)$$

Using (9) we get

$$\lim_{r\to\infty} |\sin\pi(\nu+s)\sin\pi s| = \lim_{r\to\infty} \frac{1}{4} e^{2\pi r|\sin\theta|} \qquad (c)$$

Substituting (c) and the asymptotic expansions of $\Gamma(1+\nu+s)$ and $\Gamma(s)$, (b) becomes

$$L_2 = 2\pi \lim_{r\to\infty} |e^{A_2}| \qquad (d)$$

where

$$A_2 = (\nu+2s)\log\frac{-iz}{2} + (\nu+2s+1) - 2\pi r|\sin\theta| - (\nu+s+1)\log(\nu+s+1) - (s-\tfrac{1}{2})\log s$$

which we transform to

$$A_2 = (\nu+2s)\log\frac{-ize}{2d} - (\nu+s+\tfrac{1}{2})\log\frac{\nu+s+1}{s} - 2\pi r|\sin\theta|+1 \qquad (e)$$

Taking the real part of A_2, we find

$$L_2 = 2\pi \lim_{r\to\infty} \left(\frac{e|-iz|}{2r}\right)^{\nu+2r\cos\theta} e^{2rB+1}$$

where

$$B = -\arg(-iz)\sin\theta + \theta\sin\theta - \pi|\sin\theta|$$

Because

$$B \leq 0$$

and

$$\nu+2r\cos\theta \geq 0$$

in the range $-\pi\leq 2\theta<\pi$, it is proved that

$$\lim_{r\to\infty}|L_2| = \text{finite} \qquad \text{when } R(\nu) = 0 \quad \text{and } \theta = \frac{\pi}{2}$$

$$= 0 \qquad \text{in the other cases.}$$

If $\nu = 0$ and $\theta = \frac{\pi}{2}$ or $-\frac{\pi}{2}$, L_2 may be finite but never becomes infinite.

Therefore, we can conclude that Barnes' theory by use of (10) is valid for $\nu \geqq 0$ and $|\arg(-iz)| \leqq \frac{\pi}{2}$, unless the order of the integration and the substituting $z = 0$ is exchanged. Q.E.D.

We will show in the following examples that indiscriminate use of Barnes' integral representation is dangerous.

According to the theorem stated above, we have the indentities,

$$\int_0^\infty \frac{1}{1+x^4} J_1(ax)\, J_0(bx)\, dx = \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{1+x^4} H_1^{(1)}(ax)\, J_0(bx)\, dx \qquad (11)$$

when $a>b$, and

$$\int_0^\infty \frac{1}{1+x^4} J_1(ax) J_0(bx)\, dx = \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{1+x^4} J_1(ax)\, H_0^{(1)}(bx)\, dx \qquad (12)$$

when $a>b$. The ranges of integration in these integrals are restricted on real axis; the circular contour in Figure 1 is not considered.

Substituting the respective Barnes' representations on the right hand sides of (11) and (12), and changing the order of integrations, yields the results in Example 1, as will be shown in Example 2.

Substituting the respective Barnes' representations on the left hand sides of (11) and (12), and changing the order of integrations,

does not necessarily yields the results in Example 1, as will be shown in Examples 3 and 4. The difference is caused by the pole of $H_1^{(1)}(ax)$ at $x=0$.

The correct result must be the one found by substituting the respective Barnes' representations into the right hand sides, because the integrals on the left hand sides of (11) and (12), considered as functions of a and b, are continuous at $a=b$. This condition is not satisfied by substituting the respective Barnes' representations into the left hand sides, but into the right hand sides, as shown in Example 1.

Example 2.

We shall show in the following that use of Barnes' theory to integrate the right hand sides of (11) and (12), which are expressed here as

$$I(a>b) = \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{1+x^4} H_1^{(1)}(ax)\, J_0(bx)\,dx \qquad (13)$$

and

$$I(a<b) = \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{1+x^4} J_1(ax)\, H_0^{(1)}(bx)\,dx \qquad (14)$$

yields the same result as given in Example 1 by use of the contour integral method.

Using Barnes' representation (10) and (7) for $H_1^{(1)}(ax)$ and $J_0(bx)$, respectively, and changing the order of integration, integral (13) becomes

$$I(a>b) = \frac{1}{2} - \frac{1}{2\pi^2 i}\int_{-1-\infty i}^{-1+\infty i}\Gamma(-1-s)\Gamma(-s)\left(-\frac{ia}{2}\right)^{1+2s} ds \frac{1}{2\pi i}\int_{-\infty i}^{\infty i}\frac{\Gamma(-t)\left(\frac{b}{2}\right)^{2t}}{\Gamma(t+1)} K_1\, dt + \frac{1}{a} \tag{a}$$

where

$$K_1 = \lim_{\varepsilon\to 0}\int_{\varepsilon}^{\infty}\frac{x^{1+2s+2t}}{1+x^4}dx + \lim_{\varepsilon\to 0}\int_{-\infty}^{-\varepsilon}\frac{x^{1+2s+2t}}{1+x^4}dx \tag{b}$$

The last term on the right hand side of (a) results from the existence of the pole at $x = 0$ in the integrande of (13). The residue at this pole is evaluated by drawing a semi-circle as in Figure 1.

Changing s to $s-1$ in the first integral in (a), $I(a<b)$ becomes

$$I(a>b) = \frac{1}{2}\left(-\frac{1}{2\pi^2 i}\right)\int_{-\infty i}^{\infty i}\Gamma(-s)\Gamma(-s+1)\left(\frac{ia}{2}\right)^{-1+2s} ds \frac{1}{2\pi i}\int_{-\infty i}^{\infty i}\frac{\Gamma(-t)\left(\frac{b}{2}\right)^{2t}}{\Gamma(t+1)} K_1 dt + \frac{1}{a} \tag{c}$$

where

$$K_1 = \lim_{\varepsilon\to 0}\int_{\varepsilon}^{\infty}\frac{x^{-1+2s+2t}}{1+x^4}dx + \lim_{\varepsilon\to 0}\int_{-\infty}^{-\varepsilon}\frac{x^{-1+2s+2t}}{1+x^4}dx \tag{d}$$

Changing x to $-x$ in the second integral of (d), K_1 becomes

$$K_1 = \left(1+e^{i\pi(-1+2s+2t)}\right)\int_{0}^{\infty}\frac{x^{-1+2s+2t}}{1+x^4}dx \tag{e}$$

Letting $\xi = x^4$, (e) integrates to

$$K_1 = \frac{1}{4}\left(1-e^{-2i\pi(s+t)}\right)\Gamma\left(\frac{s+t}{2}\right)\Gamma\left(1-\frac{s+t}{2}\right)$$

which transforms to

$$K_1 = \frac{\pi}{4}\left(1-e^{2i\pi(s+t)}\right)\frac{1}{\sin\frac{\pi}{2}(s+t)} \qquad (f)$$

$$= -i\pi e^{\pi i(s+t)}\cos\frac{\pi}{2}(s+t)$$

Substituting (f), (c) decomposes into products of single integrals,

$$I(a>b) = \frac{i}{2}\left(I_1 I_2 + I_2 I_4\right) + \frac{1}{a} \qquad (g)$$

where

$$I_1 = \frac{-\frac{\pi}{2}}{2\pi i}\int_{-\infty i}^{\infty i}\frac{\left(-\frac{ia}{2}\right)^{-1+2s}e^{i\pi s}}{\Gamma(s+1)}\,\frac{\Gamma(-s+1)}{\sin\frac{\pi s}{2}}\,ds$$

$$I_2 = \frac{-\frac{\pi}{2}}{2\pi i}\int_{-\infty i}^{\infty i}\frac{\left(\frac{b}{2}\right)^{2t}e^{i\pi t}}{(\Gamma(t+1))^2}\,\frac{dt}{\sin\frac{\pi t}{2}}$$

$$I_3 = \frac{\frac{\pi}{2}}{2\pi i}\int_{-\infty i}^{\infty i}\frac{\left(-\frac{ia}{2}\right)^{-1+2s}e^{i\pi s}}{\Gamma(s+1)}\,\frac{\Gamma(-s+1)}{\cos\frac{\pi s}{2}}\,ds$$

$$I_4 = \frac{-\frac{\pi}{2}}{2\pi i}\int_{-\infty i}^{\infty i}\frac{\left(\frac{b}{2}\right)^{2t}e^{i\pi t}}{(\Gamma(t+1))^2}\,\frac{dt}{\cos\frac{\pi t}{2}}$$

The integrand of I_1 has a single pole at $s = 0$ and $s = 2n+1$ and a double pole at $s = 2n$, where $n = 1, 2, \ldots$. Let

$$f(s) = \frac{\left(-\frac{ia}{2}\right)^{-1+2s}e^{i\pi s}}{\Gamma(s+1)\,\Gamma(s)} \qquad (h)$$

Then I_1 becomes

$$I_1 = -\frac{2}{ia} + \sum_{n=1}^{\infty} (-1)^n f'(2n) - \frac{\pi}{2} \sum_{n=0}^{\infty} (-1)^n f(2n+1) \qquad \text{(i)}$$

The first, second, and third terms on the right hand side of (i) are the residues at $s = 0$, $s = 2n$, and $s = 2n+1$, respectively. Substituting (h), (i) becomes

$$I_1 = -2i\,\mathrm{ker}'a$$

where

$$\mathrm{ker}'x = -\frac{1}{x} + \frac{\pi}{4} \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!(2n+1)!} \left(\frac{x}{2}\right)^{4n-1} +$$

$$+ \sum_{n=0}^{\infty} \frac{(-1)^n \left(\frac{x}{2}\right)^{4n+3}}{(2n+1)!(2n+2)!} \left\{ \log \frac{x}{2} - \frac{1}{2}\left(\psi(2n+2)+\psi(2n+3)\right) \right\}$$

The integral I_2 has a single pole at $t = 2n$, where $n = 0,1,2, \ldots$. Counting the sume of the residues, I_2 becomes

$$I_2 = \mathrm{ber}\ b$$

The integral I_3 has a single pole at $s = 2n$ and a double pole at $s = 2n+1$, where $n = 0,1,2,\ldots$. Counting the sum of the residues. I_3 becomes

$$I_3 = -2i\ \mathrm{kei}'a$$

where

$$\text{kei}'x = \frac{\pi}{4}\sum_{n=0}^{\infty} \frac{(-1)^n (\frac{x}{2})^{4n+3}}{(2n+1)!(2n+2)!} -$$

$$- \sum_{n=0}^{\infty} \frac{(-1)^n (\frac{x}{2})^{4n+1}}{(2n)!(2n+1)!} \left\{ \log\frac{x}{2} - \frac{1}{2}\left(\psi(2n+1)+\psi(2n+2)\right) \right\}$$

The integral I_4 has a single pole at $t = 2n+1$, where $n = 0,1,2,\ldots$. Counting the sum of the residues, I_4 becomes

$$I_4 = -\ \text{bei}\ b$$

Substituting the above values of I_1, I_2, I_3 and I_4 into (g), we can show that $I(a>b)$ in (13) is given by (5).

Using Barnes' representations (7) and (10) for $J_1(ax)$ and $H_o^{(1)}(bx)$, respectively, and changing the order of integration, integral (14) becomes

$$I(a<b) = \frac{1}{(2\pi i)^2} \int_{-\infty i}^{\infty i} (\Gamma(-t))^2 \left(\frac{-ib}{2}\right)^2 dt \frac{1}{2\pi t} \int_{-\infty i}^{\infty i} \frac{\Gamma(-s)(\frac{a}{2})^{1+2s}}{\Gamma(s+2)} K_2 ds \qquad (j)$$

where

$$K_2 = \int_{-\infty}^{\infty} \frac{x^{1+2s+2t}}{1+x^4}\, dx$$

Integration of K_2 yields

$$K_2 = -\ i\pi\ e^{i\pi(s+t)} \sin\frac{\pi}{2}(s+t) \qquad (k)$$

Substituting (k), (j) decomposes into products of single integrals,

$$I(a<b) = -\frac{1}{2}\left(I_1\, I_2 + I_3\, I_4\right) \qquad (l)$$

where

$$I_1 = \frac{\frac{\pi}{2}}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\left(\frac{a}{2}\right)^{1+2s} e^{i\pi s}}{\Gamma(s+2)\Gamma(s+2)\cos\frac{\pi s}{2}}\, ds$$

$$I_2 = \frac{\left(\frac{\pi}{2}\right)^2}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\left(\frac{-ib}{2}\right)^{2t} e^{i\pi t}}{\left(\Gamma(t+1)\right)^2 \sin\pi t \sin\frac{\pi t}{2}}\, dt$$

$$I_3 = \frac{\frac{-\pi}{2}}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\left(\frac{a}{2}\right)^{1+2s} e^{i\pi s}}{(\Gamma(s+1)\Gamma(s+2)\sin\frac{\pi s}{2}}\, ds$$

and

$$I_4 = \frac{\left(\frac{\pi}{2}\right)^2}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\left(\frac{-ib}{2}\right)^{2t} e^{i\pi t}}{\left(\Gamma(t+1)\right)^2 \sin\pi t \cos\frac{\pi t}{2}}\, dt$$

They integrate to

$$I_1 = \mathrm{ber}'a$$

$$I_2 = -2\,\mathrm{ker}b$$

$$I_3 = \mathrm{bei}'a$$

and

$$I_4 = 2\,\mathrm{kei}b$$

Substituting I_1, I_2, I_3, and I_4 in the above we can show that $I(a<b)$ in (14) is given by (6).

Example 3.

We shall show in the following that use of Barnes' theory to integrate the left hand sides of (11) and (12), which are expressed here as

$$J = \int_0^\infty \frac{1}{1+x^4} J_1(ax)\, J_0(bx)\, dx \tag{15}$$

yields

$$J = \text{ber}\,(\tfrac{1}{a} + \text{ker}'\, a) - \text{bei}\, b\ \text{kei}'\, a \tag{16}$$

when $a>b$

$$= \text{ber}'a\ \text{ker}b - \text{bei}'a\ \text{kei}b \tag{17}$$

when $a<b$

The result (16) for $a>b$ does not agree with (5). The result (17) for $a<b$ agrees with (6). The difference is caused by the pole of $H^{(1)}(ax)$ at $x=0$, which is not counted in (15). The meaning of the value of (16) is not yet known.

We shall use a single Barnes' representation for the product $J_1(ax)\, J_0(bx)$ in the example. Two expressions are available (Watson 2, p. 148).

$$J_1(ax)\, J_0(bx) = \frac{a}{b} \sum_{n=0}^{\infty} \frac{(-1)^n (\frac{bx}{2})^{2n+1}}{(n!)^2} F(-n,-n;2;\frac{a^2}{b^2}) \tag{a}$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n\ (\frac{ax}{2})^{2n+1}}{n!\,(n+1)!} F(-n,\ -n-1;\ 1;\ \frac{b^2}{a^2}) \tag{b}$$

where F(, ; ;) denotes a hypergeometric function. Both hypergeometric functions are polynomials of nth order.

Use of (a) yields the integral for $a<b$, as shown in the following. The integral (15) for this case will be denoted by $J(a<b)$ in anticipation.

Transform (a) to a Barnes' representation:

$$J_1(ax)J_o(bx) = \frac{a}{b}\ \frac{1}{2\pi i}\int_{-\infty i}^{\infty i} \frac{\Gamma(-s)\left(\frac{bx}{2}\right)^{2s+1}}{\Gamma(s+1)}\ F\left(-s,\ -s;\ 2;\ \frac{a^2}{b^2}\right) ds \qquad (c)$$

Substituting this into (15) and changing the order of integration, we have

$$J(a<b) = \frac{a}{b}\ \frac{1}{2\pi i}\int_{-\infty i}^{\infty} \frac{\Gamma(-s)\left(\frac{b}{2}\right)^{2s+1}}{\Gamma(s+1)}\ F\left(-s,-s;2;\frac{a^2}{b^2}\right) ds \int_o^{\infty} \frac{x^{2s+1}}{1+x^4}\ dx \qquad (d)$$

We must examine the asymptotic behavior of

$$F(-s,\ -s;\ 2;\frac{a^2}{b^2}) = \frac{1}{\Gamma(-s)\Gamma(2+s)} \int_o^1 \xi^{-s-1}\ (1-\xi)^{1+s}\ \left(1-\frac{a^2}{b^2}\xi\right)^s\ d\xi$$

For $\frac{a}{b^2} \leq 1$, we have

$$\left(1-\frac{a^2}{b^2}\xi\right)^s \leq 1$$

because we assume $R(s) \geq 0$. Therefore we have

$$F\left(-s,\ -s;\ 2;\ \frac{a^2}{b^2}\right) \leq 1 \qquad (e)$$

for any values of s when $a<b$. To prove (e), note that

$$\int_o^1 \xi^{-s-1}(1-\xi)^{1+s}\ d\xi = \Gamma\left(-s\right)\Gamma\left(2+s\right)$$

Letting $x^4 = \xi$, the second integral on the right hand side of (d) is integrated to:

$$\int_o^{\infty} \frac{x^{2s+1}}{1+x^4}\ dx = \frac{\frac{\pi}{4}}{\cos\frac{\pi s}{2}}$$

Letting $x^4 = \xi$, the second integral on the right hand side of (d) is integrated to:

$$\int_0^\infty \frac{x^{2s+1}}{1+x^4}\,dx = \frac{\frac{\pi}{4}}{\cos\frac{\pi s}{2}}$$

Thus (d) becomes

$$J(a<b) = \frac{-\frac{\pi^2}{4}}{2\pi i}\int_{-\infty i}^{\infty i} f_1(s)\,\frac{ds}{\cos\frac{\pi s}{2}\sin\pi s} \qquad (f)$$

where

$$f_1(s) = \frac{a}{b}\,\frac{(\frac{b}{2})^{2s+1}}{\Gamma^2(s+1)}\,F\left(-s,-s;2;\frac{a^2}{b^2}\right) \qquad (g)$$

For $a \leq b$ we may apply the residue theorem to integrate (f). The integrand of (f) has a single pole at $s=2n$ and a double pole at $s = 2n+1$, where $n = 0,1,2,\ldots$. Thus (f) integrates to

$$J(a<b) = \frac{\pi}{4}\sum_{n=0}^{\infty}(-1)^n f_1(2n) + \frac{1}{2}\sum_{n=0}^{\infty}(-1)^n f_1'(2n+1) \qquad (h)$$

Substituting (g), (h) becomes

$$J(a<b) = \frac{\pi}{4}K_1 + \log\frac{b}{2} - M_1 + \frac{1}{2}N_1 \qquad (i)$$

where

$$K_1 = \frac{a}{b}\sum_{n=0}^{\infty}\frac{(-1)^n(\frac{b}{2})^{4n+1}}{((2n)!)^2}\,F(-2n,-2n;\ 2;\ \frac{a^2}{b^2})$$

$$L_1 = \frac{a}{b}\sum_{n=0}^{\infty}\frac{(-1)^n(\frac{b}{2})^{4n+3}}{((2n+1)!)^2}\,F(-2n-1,-2n-1;2;\ \frac{a^2}{b^2})$$

$$M_1 = \frac{a}{b}\sum_{n=0}^{\infty}\frac{(-1)^n\left(\frac{b}{2}\right)^{4n+3}}{(2n+1)!)^2}\quad F(-2n-1,-2n-1;\ 2;\ \frac{a^2}{b^2})\psi(2n+2)$$

and

$$N_1 = \frac{a}{b}\sum_{n=0}^{\infty}\frac{(-1)^n\left(\frac{b}{2}\right)^{4n+3}}{((2n+1)!)^2}\quad \left[\frac{\partial}{\partial s}F(-s,-s;\ 2;\ \frac{a^2}{b^2})\right]_{s=2n+1}$$

Expressing $F(-2n,\ 2n;\ 2;\ \frac{a^2}{b^2})$ as a polynomial, K_1 becomes

$$K_1 = \sum_{n=0}^{\infty}(-1)^n(\tfrac{b}{2})^{4n+1}\sum_{k=0}^{2n}\frac{1}{((2n-k)!)^2k!(k+1)!}\left(\frac{a}{b}\right)^{2k+1}$$

Dividing the second series into two series with regard to even integers and odd integers, K_1 becomes

$$K_1 = \sum_{n=0}^{\infty}(-1)^n(\tfrac{b}{2})^{4n+1}\sum_{p=0}^{n}\frac{1}{(2n-2p)!)^2(2p)!(2p+1)!}\left(\frac{a}{b}\right)^{4p+1}$$

$$+\sum_{n=1}^{\infty}(-1)^n(\tfrac{b}{2})^{4n+1}\sum_{p=1}^{n}\frac{1}{(2n-2p+1)!(2p-1)!(2p)!}\left(\frac{a}{b}\right)^{4p-1}$$

Letting $n = p+q$, K_1 becomes

$$K_1 = \sum_{p=0}^{\infty}\frac{(-1)^p}{(2p)!(2p+1)!}\left(\frac{a}{2}\right)^{4p+1}\sum_{q=0}^{\infty}\frac{(-1)^q}{((2q)!)^2}\left(\frac{b}{2}\right)^{4q}+$$

$$+\sum_{p=1}^{\infty}\frac{(-1)^p}{(2p)!(2p-1)!}\left(\frac{a}{2}\right)^{4p-1}\sum_{q=0}^{\infty}\frac{(-1)^q}{((2q+1)!)^2}\left(\frac{b}{2}\right)^{4q+2}$$

Thus we get

$$K_1 = \text{bei}'a \ker b + \text{ber}'a \ \text{bei} b \qquad (j)$$

Expressing $F(-2n-1, -2n-1; 2; \frac{a^2}{b^2})$ as polynomial, L_1 becomes

$$L_1 = \sum_{n=0}^{\infty} (-1)^n (\frac{b}{2})^{4n+3} \sum_{k=0}^{2n+1} \frac{1}{((2+1-k)!)^2 k!(k+1)!} (\frac{a}{b})^{2k+1}$$

Dividing the second series into two series with regard to even integers and odd integers,

$$L_1 = \sum_{n=0}^{\infty} (-1)^n (\frac{b}{2})^{4n+3} \sum_{p=0}^{n} \frac{1}{((2n+1-2p)!)^2 (2p)!(2p+1)!} (\frac{a}{b})^{4p+1} +$$

$$+ \sum_{n=0}^{\infty} (-1)^n (\frac{b}{2})^{4n+3} \sum_{p=0}^{n} \frac{1}{((2n-2p)!)^2 (2p+1)!(2p+2)!} (\frac{a}{b})^{4p+3}$$

Letting $n = p+q$, L_1 becomes

$$L_1 = \sum_{p=0}^{\infty} \frac{(-1)^p}{(2p)!(2p+1)!} (\frac{a}{2})^{4p+1} \sum_{q=0}^{\infty} \frac{(-1)^q}{((2q+1)!)^2} (\frac{b}{2})^{4q+2} +$$

$$+ \sum_{p=0}^{\infty} \frac{(-1)^p}{(2p+1)!(2p+2)!} (\frac{a}{2})^{4p+3} \sum_{q=0}^{\infty} \frac{(-1)^q}{((2q)!)^2} (\frac{b}{2})^{4q}$$

Thus we get

$$L_1 = \text{bei}'a\ \text{bei}b - \text{ber}'a\ \text{ber}b \qquad (k)$$

Next we compute

$$O_1 = \left[\frac{\partial}{\partial s} F(-s, -s; 2; \frac{a^2}{b^2})\right]_{s=2n+1}$$

which is

$$O_1 = \left[\frac{\partial}{\partial s} \sum_{k=0}^{\infty} \frac{((-s)_k)^2}{k!\ (k+1)!} (\frac{a}{b})^{2k}\right]_{s=2n+1}$$

By differentiation,

$$O_1 = 2 \sum_{k=0}^{\infty} \frac{((-2n-1)_k)^2}{k!(k+1)!} (\frac{a}{b})^{2k} \sum_{h=0}^{k-1} \frac{1}{2n+1-h}$$

Thus we find:

$$-M_1 + \frac{1}{2} N_1$$

$$= \sum_{n=0}^{\infty} (-1)^n (\frac{b}{2})^{4n+3} \sum_{k=0}^{2n+1} \frac{1}{((2n+1-k)!)^2 k!(k+1)!} (\frac{a}{b})^{2k+1} P_{n,k}$$

where

$$P_{n,k} = -\psi(2n+2) + \sum_{h=0}^{k-1} \frac{1}{2n+1-h}$$

which transforms to

$$P_{n,k} = -\psi(2n+2-k)$$

Thus we have

$$- M_1 + \frac{1}{2} N_1 = - \sum_{n=0}^{\infty} (-1)^n (\frac{b}{2})^{4n+3} \sum_{k=0}^{2n+1} \frac{\psi(2n+2-k)}{((2n+1-k)!)^2 k!(k+1)!} (\frac{a}{b})^{2k+1}$$

Dividing the second series into two series with regard to even integers and odd integers,

$$-M_1 + \frac{1}{2} N_1 = - \sum_{n=0}^{\infty} (-1)^n (\frac{b}{2})^{4n+3} \sum_{p=0}^{n} \frac{\psi(2n+-2-2p)}{((2n+1-2p)!)^2 (2p)!(2p+1)!} (\frac{a}{b})^{4p+1}$$

$$- \sum_{n=0}^{\infty} (-1)^n (\frac{b}{2})^{4n+3} \sum_{p=0}^{n} \frac{\psi(2n+1-2p)}{((2n-2p)!^2 (2p+1)!(2p+2)!} (\frac{a}{b})^{4p+3}$$

Letting $n = p+q$, $-M_1 + \frac{1}{2} N_1$ becomes

$$-M_1 + \frac{1}{2} N_1 = - \sum_{p=0}^{\infty} \frac{(-1)^p}{(2p)!(2p+1)!} (\frac{a}{2})^{4p+1} \sum_{q=0}^{\infty} \frac{(-1)^q \psi(2q+2)}{((2q+1)!)^2} (\frac{b}{2})^{4q+2}$$

$$- \sum_{p=0}^{\infty} \frac{(-1)^p}{(2p+1)!(2p+2)!} (\frac{a}{2})^{4p+3} \sum_{q=0}^{\infty} \frac{(-1)^q \psi(2q+1)}{((2q)!^2} (\frac{b}{2})^{4q}$$

$$= - \text{bei}'a \sum_{q=0}^{\infty} (\text{bei}b)_{4q+2} \psi(2q+2) +$$

$$+ \text{ber}'_a \sum_{q=0}^{\infty} (\text{ber}b)_{4q} \psi(2q+1) \qquad (1)$$

where $(\text{bei}b)_{4q+2}$ and $(\text{ber}b)_{4q}$ denote the $(4q+2)$th and $(4q)$th order terms of the series of bei$b$ and ber$b$ respectively.

Substituting (j), (k), and (l) into (i) we find that $J(a<b)$ is equal to (17).

Use of expression (b) of $J_1(ax)\,J_o(bx)$ to integrate (15) yields the integral for $a>b$, which will be denoted by $J(a>b)$ in anticipation.

Transform (b) to a Barnes' representation:

$$J_1(ax)J_o(bx) = \frac{1}{2\pi i}\int_{-\infty i}^{\infty i} \frac{\Gamma(-s)(\frac{ax}{2})^{2s+1}}{\Gamma(s+2)}\ F(-s,-s-1;1;\frac{b^2}{a^2})\ ds \qquad (m)$$

Substituting (m) into (15), and changing the order of integration, we have

$$J(a>b) = \frac{1}{2\pi i}\int_{-\infty i}^{\infty i} \frac{\Gamma(-s)(\frac{a}{2})^{2s+1}}{\Gamma(s+2)}\ F(-s,-s-1;1;\frac{b^2}{a^2})\ ds \int_o^\infty \frac{x^{2s+1}}{1+x^4}\ dx \qquad (n)$$

It can be proved that

$$\left|F(-s,\ -s-1;1;\frac{b^2}{a^2})\right| \leq 1$$

for $a \geq b$. Following the similar procedure to the previous case, we get

$$J(a>b) = \frac{-\frac{\pi^2}{4}}{2\pi i}\int_{-\infty i}^{\infty i} f_2(s)\ \frac{ds}{\cos\frac{\pi s}{2}\sin\pi s}$$

where

$$f_2(s) = \frac{(\frac{a}{2})^{2s+1}}{\Gamma(s+1)\Gamma(s+2)}\ F(-s,-s-1;1;\frac{b^2}{a^2})$$

It integrates to

$$J(a>b) = \frac{\pi}{4} K_2 + L_2 \log \frac{a}{2} - M_2 + \frac{1}{2} N_2 \qquad (o)$$

where

$$K_2 = \sum_{n=o}^{\infty} \frac{(-1)^n (\frac{a}{2})^{4n+1}}{(2n)!(2n+1)!} \quad F(-2n, -2n-1; 1; \frac{b^2}{a^2})$$

$$L_2 = \sum_{n=o}^{\infty} \frac{(-1)^n (\frac{a}{2})^{4n+3}}{(2n+1)!(2n+2)!} \quad F(-2n-1, -2n-2; 1; \frac{b^2}{a^2})$$

$$M_2 = \frac{1}{2} \sum_{n=o}^{\infty} \frac{(-1)^n (\frac{a}{2})^{4n+3}}{(2n=L)!(2n+2)!} \quad F(-2n-1, -2n-2; 1; \frac{b^2}{a^2})(\psi(2n+2)+\psi(2n+3))$$

and

$$N_2 = \sum_{n=o}^{\infty} \frac{(-1)^n (\frac{a}{2})^{4n+3}}{(2n+1)!(2n+2)!} \left[\frac{\partial}{\partial s} \quad F(-s, -s-1; 1; \frac{b^2}{a^2})\right]_{s=2n+1}$$

Following the similar procedure to the previous case we find

$$K_2 = \text{bei}'a \ \text{ber}b + \text{ber}'a \ \text{ber}b$$

$$L_2 = \text{bei}'a \ \text{bei}b - \text{ber}'a \ \text{ber}b$$

$$\left[\frac{\partial}{\partial s} F(-s, -s-1; 1; \frac{b^2}{a^2})\right]_{s=2n+1}$$

$$= \sum_{k=o}^{2n+1} \frac{(2n+1)!(2n+2)! \ (\frac{b}{a})^{2k}}{(2n+1-k)!(2n+2-k)! \ (k!)^2} \sum_{h=o}^{k-1} \left(\frac{1}{2n+1-h} + \frac{1}{2n+2-h}\right)$$

$$-M_2 + \frac{1}{2} N_2$$

$$= \operatorname{ber} b \sum_{q=0}^{\infty} \left(\operatorname{ber}' a\right)_{4q+3} (\psi(2q+2) + \psi(2q+3)) -$$

$$- \operatorname{bei} b \sum_{q=0}^{\infty} (\operatorname{bei}' a)_{4q+1} (\psi(2q+1) + \psi(2q+2))$$

Substituting these formulas into (o), we find that $J(a>b)$ is given by (16)

Example 4.

We shall use in the following two Barnes' representations to integrate (15). Using (7) for both $J_1(ax)$ and $J_0(bx)$ and changing the order of integration, (15) becomes

$$J = \frac{1}{(2\pi i)^2} \int_{-\infty i}^{\infty i} \frac{\Gamma(-s)\left(\frac{a}{2}\right)^{1+2s}}{\Gamma(s+2)} ds \int_{-\infty i}^{\infty i} \frac{\Gamma(-t)\left(\frac{b}{2}\right)^{1+2t}}{\Gamma(t+1)} dt \int_0^{\infty} \frac{x^{1+2s+2t}}{1+x^4} dx \qquad (a)$$

Two cases occur in accordance with the order of integration. We shall show that, when integration with regard to t, or s, is performed first, J yields (16), or (17), respectively. These integrations will be denoted by $J(a>b)$ and $J(a<b)$, respectively, as in the previous example.

To prove this, let

$$s + t = u \qquad (b)$$

in (a). Assuming that integration with regard to t is performed first, we transform (a) to

$$J = \frac{1}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\Gamma(-u)\left(\frac{a}{2}\right)^{1+2u}}{\Gamma(u+2)} F \, du \int_0^{\infty} \frac{x^{1+2u}}{1+x^4} \, dx \qquad (c)$$

where

$$F = \frac{1}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\Gamma(t-u)}{\Gamma(-u)} \frac{\Gamma(u+2)}{\Gamma(u-t+2)} \frac{\Gamma(-t)}{\Gamma(t+1)} \left(\frac{b}{a}\right)^{2t} dt \qquad (d)$$

We shall show that F in (d) transforms to

$$F = F\left(-u, -u-1; 1; \frac{b^2}{a^2}\right) \qquad (e)$$

To prove (e), apply (8) to $\Gamma(u+2)$ and $\Gamma(u - t+2)$ in (d), and we have

$$F = \frac{1}{2\pi i} \frac{-\pi}{\Gamma(-u)\Gamma((-u-1)} \int_{-\infty i}^{\infty i} f(t) \frac{dt}{\sin \pi t} \qquad (f)$$

where

$$f(t) = \frac{\Gamma(t-u)\Gamma(t-u-1)}{\Gamma^2(t+1)} \left(\frac{b}{a}\right)^{2t} \frac{\sin\pi(u-t+2)}{\sin\pi(u+2)}$$

On integration (f) becomes (e). Then J in (c) is equal to $J(a>b)$. The other case can be proved similarly.

In the following we shall carry out the integration of $J(a>b)$. In this case we first perform the integration

$$M = -\frac{\pi^2}{4} \frac{1}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\left(\frac{b}{2}\right)^{2t}}{\Gamma^2(t+1)\cos\frac{\pi(s+t)}{2}} \frac{dt}{\sin \pi t} \qquad (g)$$

which is found by substituting

$$\int_0^\infty \frac{x^{1+2s+2t}}{1+x^4}\,dx = \frac{\frac{\pi}{4}}{\cos\frac{\pi(s+t)}{2}}$$

into (a). Because s is imaginary, the integrand of (g) has a single pole at $t = n$, where $n = 0,1,2, \ldots$. Thus (g) becomes

$$M = \frac{\pi}{4}\sum_{n=0}^{\infty} \frac{(-1)^n(\frac{b}{2})^{2n}}{(n!)^2} \frac{1}{\cos(\frac{n\pi}{4}+\frac{\pi s}{2})}$$

Dividing the above series into the series of even integers and odd integers, M becomes

$$M = \frac{\pi}{4}\left(\mathrm{ber}b \frac{1}{\cos\frac{\pi s}{2}} + \mathrm{bei}b \frac{1}{\sin\frac{\pi s}{2}}\right)$$

Substituting this result, $J(a>b)$ becomes

$$J(a>b) = \frac{\pi^2}{4}(N_1\mathrm{ber}b + N_2\mathrm{bei}b)$$

where

$$N_1 = -\frac{1}{2\pi i}\int_{-\infty i}^{\infty i} \frac{(\frac{a}{2})^{1+2s}}{\Gamma(s+1)\Gamma(s+2)} \frac{ds}{\sin\pi s \cos\frac{\pi s}{2}}$$

and

$$N_2 = -\frac{1}{2\pi i}\int_{-\infty i}^{\infty i} \frac{(\frac{a}{2})^{1+2s}}{\Gamma(s+1)\Gamma(s+2)} \frac{ds}{\sin\pi s \sin\frac{\pi s}{2}}$$

The integrand of N_1 has a single pole at $s = 2n$ and a double pole at $s = 2n+1$, where $n = 0,1,2, \ldots$. Counting the sum of the residues, N_1 becomes

$$N_1 = \frac{4}{\pi^2} \left(\ker' a + \frac{1}{a}\right)$$

The integrand of N_2 has a double pole at $s = 2n$ and a single pole at $s = 2n+1$, where $n = 0,1,2, \ldots$. Counting the sum of the residues, N_2 becomes

$$N_2 = -\frac{4}{\pi^2} \operatorname{kei}' a$$

Thus we can conclude that $J(a>b)$ is given by (16).

Similarly we can find that $J(a<b)$ is given by (17).

Example 5.

We shall show in this example complicated variants of (1) whose integration procedure is not obvious.

The most complicated integrals we encountered in the study of viscoelastic deformation of the floating ice plate (Ref. 1) are:

$$I_1 = \int_0^\infty \frac{\tau-\alpha_1}{\sqrt{D(x)}} \left(1-e^{-\alpha_1(x)t}\right) J_1(ax)\, J_0(bx) x^4\, dx \tag{18}$$

and

$$I_2 = \int_0^\infty \frac{\tau-\alpha_2}{\sqrt{D(x)}} \left(1-e^{-\alpha_2(x)t}\right) J_1(ax)\, J_0(bx) x^4\, dx \tag{19}$$

Here viscoelastic constants E_1, E_2, η_1, and η_2 are used to define

$$\tau = \frac{\eta_2}{\eta_1}\left(1+\frac{E_2}{E_1}\right)$$

$$E = \frac{E_2}{E_1 + E_2}$$

$$D(x) = [\tau(1+x^4)-1]^2 + 4\tau(1-E)$$

$$\alpha_1(x) = \frac{\tau x^4 + 1 + \tau - \sqrt{D(x)}}{2(E+x^4)}$$

$$\alpha_2(x) = \frac{\tau x^4 + 1 + \tau + \sqrt{D(x)}}{2(E+x^4)}$$

and t is time. Integral I_2 has essential singularities at the roots of $x^4 = -E$, because

$$\lim_{x^4 \to -E} \alpha_2(x) = \infty$$

But I_1 does not, because

$$\lim_{x^4 \to -E} \alpha_1(x) = \frac{\tau}{1+\tau(1-E)}$$

Both integrals simplify, as shown in Reference (1), to

$$I_1 = \frac{1}{4} \int_0^{\lambda_1 t} \frac{1-e^{-x}}{x} J_1(af(x))\, J_0(bf(x))\, f(x)\, dx \qquad (20)$$

and

$$I_2 = \frac{1}{4} \int_{\tau t}^{\lambda_2 t} \frac{1-e^{-x}}{x} J_1(af(x)) J_0(bf(x)\, f(x)\, dx \qquad (21)$$

where

$$\lambda_1 = \frac{1}{2E}\left[1 + \tau - \sqrt{(1+\tau)^2 - 4\tau E}\right]$$

$$\lambda_2 = \frac{1}{2E}\left[1 + \tau + \sqrt{(1+\tau)^2 - 4\tau E}\right]$$

and

$$f(x) = \left(E \cdot \frac{\lambda_2 t - x}{\tau t - x} \frac{\lambda_1 t - x}{x}\right)^{\frac{1}{4}}$$

Parameters λ_1, λ_2 and τ are in the range

$$0 \leq \lambda_1 \leq \tau \leq \lambda_2$$

Integrals I_1 and I_2, as shown by (20) and (21), respectively, are continuous at $a=b$ --- a condition that must be satisfied by the results of the integrations. However, it is not clear whether the direct substitution of the Barnes' representations of (7) for $J_1(af(x))$ and $J_o(bf(x))$ does or does not yield the results satisfying this condition. Change of $J_1(af(x))$, or $J_o(bf(x))$, to respective Hankel functions may be achieved, not by simply extending x to $-x$ but by choosing a complicated contour; I am not sure, however, if I should dare to do this. I am in trouble!

REFERENCES

1. Takagi, S. and D. Nevel, Second Report on the Mathematical Study of the Viscoelastic Deformation of a Floating Ice Sheet Under a Circular Load. Technical Report, U.S. Army Cold Regions Research and Engineering Laboratory, Hanover, NH, October, 1974.

2. Watson, G.N., A Treatise on the Bessel Functions. Cambridge at the University Press, 1962.

SINGULAR PERTURBATIONS IN HEAT CONDUCTION AND DIFFUSION PROBLEMS

John F. Polk
Fluid Mechanics Branch
Applied Mathematics and Sciences Laboratory
U.S. Army Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland 21005

ABSTRACT. In one-dimensional problems of diffusion or heat conduction where discontinuities or steep gradients occur in the initial or boundary conditions a singular perturbation analysis can give accurate estimates of the solution when numerical methods prove inefficient or inadequate. In fact the discontinuities can be exploited in the singular perturbation analysis to obtain an asymptotic series representation of the solution.

Several different problems of increasing complexity can be explicitly solved when the boundary and initial data are given in piecewise polynomial form: a.) infinite region or pure initial value problem b.) semi-infinite region and c.) finite region.

The approximate methods also apply to the case when no discontinuities occur in the prescribed data or its derivatives.

1. INTRODUCTION. It is frequently stated in regard to heat conduction and diffusion problems that "discontinuities are immediately damped out". In some problems arising in engineering however this view is too over-simplified to be realistic because the damping out process itself is the heart of the problem. Typically, such problems exhibit a behavior usually referred to as "very steep gradients" which present severe difficulties to attempted solution by numerical techniques; viz., very small mesh sizes and excessive roundoff errors. We shall examine in this paper how such problems are most suitably handled by a singular perturbation analysis. The analysis will show how very accurate estimates of the solution can be obtained with just a few easily calculated terms.

We first wish to give an example of the type of problem we have in mind, namely, heat conduction in rifle barrels. We assume circular symmetry and independence of the axial coordinate. The phenomenon is then described by the heat conduction equation in radial coordinates:

$$u_t = a[u_{rr} + \frac{1}{r} u_r]$$

The radial coordinate varies from r_o = interior radius of the barrel to r_1 = exterior radius. The initial and boundary (radiation) conditions are

$$u(r,0) = \text{ambient temperature}$$

$$u(r_0,t) + h_0 u_r(r_0,t) = \text{propellant gas temperatures at time } t$$

$$u(r_1,t) + h_1 u_r(r_1,t) = \text{ambient temperature}$$

where h_0 and h_1 are heat transfer coefficients. The coefficient of thermal diffusivity, a, for mild steel is 0.12 cm^2/sec. A typical length scale in this problem is $r_1 - r_0$ which is on the order of 1 cm. A typical time scale is the duration of the phenomenon which is on the order of 1 millisecond. In non-dimensional form then the coefficient a is on the order of 10^{-4}. Although one usually thinks of metal as being a good conductor this problem is one in which the conductivity may be considered as very poor due to the short duration of the heat pulse.

An analysis of the behavior of solutions of equations of the general form

$$u_t = \varepsilon[a(x)u_{xx} + b(x)u_x + c(x)u]$$

has been undertaken by the author and explicit formulas for such problems have been obtained through singular perturbation techniques. However, in the present paper we shall deal only with the diffusion equation

$$u_t = \varepsilon u_{xx}$$

due to space limitations. The behavior for this case is simpler than the general case but typical. At first only the Cauchy problem (infinite rod) will be analyzed. Later it will be shown how extensions to mixed initial-boundary problems are easily accomplished.

There has been some treatment of this type of problem in the literature - see, for example references [1]-[4]. All of these only develop the first order approximations however and require more stringent conditions on the data than we have found necessary. In particular the case where discontinuities arise between the boundary data and the initial data has not been analyzed. The present treatment is very direct in its approach and all of the approximations will be explicitly obtained. The precise effect of discontinuities in the data and its derivatives will be clear from the asymptotic representations obtained.

The results presented here are only partial and give an indication of the general approach. A full development for the more general case will be available in the author's thesis [5] which is currently being completed at the University of Delaware.

2. Cauchy Problem for the Diffusion Equation. Consider the following problem in the region $-\infty < x < \infty$, $0 \le t < T$:

(1) $u_t = \varepsilon u_{xx}$ for $t > 0$

(2) $\lim_{t\to 0} u(x,t) = \phi(x)$ wherever $\phi(x)$ is continuous

(3) $|u(x,t)| \le M \exp[\alpha x^2]$ for $0 \le t < T$ and $-\infty < x < \infty$

where M and α are positive constants.

The existence and uniqueness of the solutions of this type of problem are thoroughly discussed by A. Friedman in [6]. In the present discussion we shall be more concerned with the computational aspects of the problem. However, in passing, we should make a few remarks about those more fundamental questions: a.) The existence of a solution is guaranteed provided $4\varepsilon\alpha T < 1$. b.) Condition (3) is sufficient to guarantee uniqueness. It is much weaker than the usual boundedness condition specified for such problems from physical reasoning. c.) Condition (2) should be replaced by a more general requirement if $\phi(x)$ is only locally integrable. However we will consider only $\phi(x)$ which are continuous except at certain isolated points where it has well defined jumps. In such cases condition (2) is sufficient. d.) Condition (3) applies in particular to $\phi(x) = u(x,0)$.

The solution of problem (1), (2), (3) can be written in the well known integral representation form

$$u(x,t) = \int_{-\infty}^{\infty} F(x-y,\varepsilon t)\phi(y)dy$$

$$\text{where } F(x,t) = \frac{1}{\sqrt{4\pi t}} \exp\,[-x^2/4t].$$

This integral is not always easy to evaluate explicitly or even numerically so we proceed with a singular perturbation analysis to obtain easily calculated solutions. We wish to emphasize here that the singular perturbation technique can be extended to more general equations where the fundamental solution is not known a priori.

3. The Functions H_n, H_n^* and v_n. In this section we introduce certain functions which will be convenient later in the discussion.

Define the initial value functions $h_n(x)$ and h_n^* for $n = 0,1,\ldots$

$$h_n(x) = \begin{cases} x^n/n! & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases}$$

and

$$h_n^*(x) = h_n(-x)$$

The functions $H_n(x,t)$ and $H_n^*(x,t)$ are then defined as the (unique) solutions of $u_t = u_{xx}$ which satisfy the growth condition (3) and the respective initial conditions

$$H_n(x,0) = h_n(x)$$

$$H_n^*(x,0) = h_n^*(x)$$

Let the two auxiliary functions $E(x,t)$ and $F(x,t)$ be

$$E(x,t) = 1/2 \text{ erfc } (-x/\sqrt{4\pi t})$$

$$F(x,t) = \frac{1}{\sqrt{4\pi t}} \exp [-x^2/4t]$$

where

$$\text{erfc}(y) = \frac{2}{\sqrt{\pi}} \int_y^\infty \exp[-z^2]dz$$

It is easy to show by induction that H_n satisfies the recursive formula

$$H_o = E$$

$$H_1 = xE + 2tF$$

$$H_n = \frac{1}{n} (xH_{n-1} + 2tH_{n-2}) \quad \text{for } n \geq 2$$

Another useful formula for H_n is

$$H_n = \frac{1}{n!} [v_n E + 2u_n F]$$

where u_n and v_n are polynomials in x and t defined by the recursive formulas

$$u_o = 0 \qquad v_o = 1$$

$$u_1 = t \qquad v_1 = x$$

$$v_n = x\, v_{n-1} + 2(n-1)t\, v_{n-2}$$

$$u_n = x\, u_{n-1} + 2(n-1)t\, u_{n-2}$$

Similarly, define E^* and F^* by

$$E^*(x,t) = E(-x,t)$$

$$F^*(x,t) = -F(x,t)$$

then we can show

$$H_o^* = E^*$$

$$H_1^* = xE^* + 2tF^*$$

$$H_n^* = \frac{1}{n} [xH_{n-1}^* + 2tH_{n-2}^*]$$

and

$$H_n^* = \frac{1}{n!} [v_n E^* + 2u_n F^*]$$

The proof of all of these is routine so we omit it. The importance of the recursive relations is that they reduce all calculations to the evaluation of the well-known functions $\exp[-x^2]$ and $\text{erfc}(x)$.

We note here that the polynomials $v_n(x,t)$ coincide with the heat polynomials discussed by Rosenbloom and Widder in [7]. Because $v_n(x,t)$ is a solution of the diffusion equation $u_t = u_{xx}$ and has initial values $v_n(x,0) = x^n$ then

$$v_n(x,t) = n!\ [H_n(x,t) + H_n^*(x,t)] \tag{5}$$

Let $L\phi = \phi_{xx}$. Then another convenient representation is

$$v_n(x,t) = \sum_{k=0}^{n'} \frac{L^k(x^n)\ t^k}{k!} \tag{6}$$

where n' is the smallest integer not less than $n/2$. Note that n' can be replaced by any larger integer since $L^k(x^n) = 0$ for any $k > n/2$.

4. Outer Solution. The solution of (1), (2) and (3) is assumed to have the asymptotic representation

$$u(x,t) = \delta_0(\varepsilon)\ u_0 + \delta_1(\varepsilon)u_1 + \cdots\cdots\cdots$$

where

$$\delta_0(\varepsilon) = 1$$

and

$$\delta_{k+1}(\varepsilon) = o(\delta_k(\varepsilon)) \quad \text{as } \varepsilon \to 0^+ \text{ for } k = 0,1,2,\ldots$$

By standard procedures one easily can show that the only reasonable choice for the asymptotic sequence is $\delta_k(\varepsilon) = \varepsilon^k$. The equations for the u_k are then obtained by substituting into (1)

$$(u_0)_t = 0$$

$$(u_k)_t = (u_{k-1})_{xx}, \qquad (k \geq 1)$$

with initial conditions

$$u_0(x,0) = \phi(x)$$

$$u_k(x,0) = 0 \qquad (k \geq 1)$$

These are easily solved when $\phi(x)$ is sufficiently differentiable,

$$u_k(x,t) = \frac{t^k L^k\ \phi\ (x)}{k!} \qquad (k \geq 0)$$

The n-term outer expansion of solutions to (1), (2), (3) is therefore given by

$$u(x,t) = \sum_{k=0}^{n} \frac{(\varepsilon t)^k\ L^k\ \phi(x)}{k!} + 0((\varepsilon t)^{n+1}) \tag{7}$$

where L^k is the operator L applied k times.

This representation is very advantageous for computations since it reduces the problem of finding u at a point (x_0,t) to the evaluation of ϕ and its derivatives at $x = x_0$. The error is of order $(\sqrt{\varepsilon t})^{2n+1}$ provided ϕ has 2n+1 continuous derivatives in a neighborhood $(x_0 - h, x_0 + h)$ where $h >> \sqrt{\varepsilon t}$. This claim is not difficult to prove using the integral representation for u(x,t) and growth properties of F(x,t). We shall not include a proof here.

In passing we note that for more general operators such as $Lu = a(x)\, u_{xx} + b(x)\, u_x + c(x)u$ the outer solution of $u_t = \varepsilon Lu$ is given by the exact same formula, (7). In fact, this representation is a formal solution if $n = \infty$ and is a true solution whenever the series is defined and convergent for all x.

5. Interior Layer. In the usual singular perturbation problem the outer solution fails to satisfy some of the data specified in the original problem and an inner solution is derived to correct any discrepancies. In the present case however the outer solution satisfies the prescribed data, given by (2), so an inner solution is not required for the usual reason. This is a rather unique feature of the pure initial value problem under consideration and should not be expected in general. In contrast, if we were concerned with a mixed BVP-IVP problem then the usual difficulties would arise.

There is, however, another source of singularities in the behavior of a perturbation solution, namely the presence of discontinuities in the prescribed data and its derivatives. It is apparent that this type of difficulty occurs with the representation (7) since any discontinuity in $\phi(x)$ or its derivatives are propagated into the solution domain along the sub-characteristics x = constant. Solutions to the diffusion equation are known to be infinitely differentiable except at the boundaries but (7) does not have that property. We are therefore led to the need for "interior layers" to correct the outer solution in the neighborhood of sub-characteristics. In general there can be many points at which $\phi(x)$ or its derivatives up to a given order do not exist. Our analysis will assume only one such point, namely $x = x_0$, but the results are easy to extend to the more general situation because of the linearity of solutions of (1).

There is another motive for studying the effect of discontinuities in the initial data for the Cauchy problem which is really the more important reason. In Section 8 we shall show how mixed boundary-initial value problems can be transformed into Cauchy type problems with discontinuous data. The results we develop now will thus be directly applicable to that case and will lead to a proper understanding of the influence of the boundary data on the solution.

To proceed with an analysis of the Cauchy problem we assume that is some neighborhood $(x_o - h, x_o + h)$ $\phi(x)$ satisfies jump conditions in the form

$$(8) \qquad \phi(x) = \begin{cases} \sum_{k=0}^{2n} (a_k/k!)\,(x - x_o)^k + R_{2n} & \text{for } x > 0 \\ \sum_{k=0}^{2n} (b_k/k!)\,(x - x_o)^k + R^*_{2n} & \text{for } x < 0 \end{cases}$$

where the Taylor remainders R_n and R^*_n are $O(x^{2n+1})$ as $x \to 0$.

To understand the behavior of $u(x,t)$ near to the sub-characteristic $x = 0$ we introduce an inner variable of the form

$$\tilde{x} = (x - x_o)/\sqrt{\varepsilon}$$

and an inner solution

$$(9) \qquad U(\tilde{x},t) = U(\sqrt{\varepsilon}\,\tilde{x} + x_o,\ t)$$

which is assumed to have an asymptotic form

$$(10) \qquad U(\tilde{x},t) = U_o(\tilde{x},t) + \delta_1(\varepsilon)\,U_1(\tilde{x},t) + \cdots\cdots$$

where $\delta_1(\varepsilon) = o(1)$ and

$$\delta_{k+1}(\varepsilon) = o(\delta_k(\varepsilon)) \qquad k \geq 1$$

Rewriting equation (1) in terms of $\tilde{x}$ gives the equation for U as

$$(11) \qquad U_t = \varepsilon\left(\frac{1}{\sqrt{\varepsilon}}\right)^2 U_{\tilde{x}\tilde{x}}$$

Substitution of (10) into (11) leads to the following equation for U_o

$$(12) \qquad (U_o)_t = (U_o)_{\tilde{x}\tilde{x}}$$

Initial conditions for $U(\tilde{x},0)$ follow from those for $u(x,t)$:

$$U(\tilde{x},0) = u(\sqrt{\varepsilon}\tilde{x},0) = \phi(\sqrt{\varepsilon}\tilde{x})$$

then from (8)

$$(13)\qquad U(\tilde{x},o) = \begin{cases} \sum_{k=0}^{2n} (a_k/k!)\ (\sqrt{\varepsilon}\tilde{x})^k + R_{2n} & \text{for } \tilde{x} > 0 \\ \sum_{k=0}^{2n} (b_k/k!)\ (\sqrt{\varepsilon}\tilde{x})^k + R^*_{2n} & \text{for } \tilde{x} < 0 \end{cases}$$

This indicates that the proper choice for the asymptotic sequence in $\delta_k(\varepsilon) = (\sqrt{\varepsilon})^k$ and the asymptotic series for U is

$$U(\tilde{x},t) = U_o(\tilde{x},t) + \sqrt{\varepsilon}U_1\ (\tilde{x},t) + \varepsilon U_2(\tilde{x},t) + \ldots\ldots$$

Substituting this expression into (11) and comparing with (13) we obtain equations and initial conditions for U_k:

$$(U_k)_t = (U_k)_{\tilde{x}\tilde{x}}$$

$$U_k(x,0) = \begin{cases} (a_k/k!)\ \tilde{x}^k & \text{for } \tilde{x} > 0 \\ (b_k/k!)\ \tilde{x}^k & \text{for } \tilde{x} < 0 \end{cases}$$

Solving for U_k in terms of H_k and H_k^* we have:

$$U_k(\tilde{x},t) = a_k\ H_k(\tilde{x},t) + b_k\ H_k^*\ (\tilde{x},t)$$

and

$$(14)\qquad U(\tilde{x},t) = \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k\ [a_k\ H_k(\tilde{x},t) + b_k\ H_k^*\ (\tilde{x},t)]$$

$$+\ O\ ((\sqrt{\varepsilon})^{2n+1})$$

Relating u and U by (9) we then set $\bar{x} = x - x_o$ and obtain

$$u(x,t) = U\ (\bar{x}/\sqrt{\varepsilon},\ t)$$

$$(15)\qquad = \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k\ [a_k\ H_k(\bar{x}/\sqrt{\varepsilon},\ t) + b_k\ H_k^*\ (\bar{x}/\sqrt{\varepsilon},t)] +$$

$$+\ O(\sqrt{\varepsilon})^{2n+1}) \text{ as } \varepsilon \to 0.$$

This representation has been derived only in a formal way but it can be rigorously justified for $|x - x_o| < h$. We now wish to indicate how (15) can be written in the form of the "outer solution" (7) plus a correction term (which we call an "interior layer") to account for the discontinuities in $\phi(x)$ or its derivatives at $x = x_o$. Letting $d_k = a_k - b_k$ (15) can be rewritten

$$u(x,t) = \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k a_k [H_k(\bar{x}/\sqrt{\varepsilon}, t) + H_k^*(\bar{x}/\sqrt{\varepsilon}, t)]$$

$$- \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k d_k H_k^*(\bar{x}/\sqrt{\varepsilon}, t) + O((\sqrt{\varepsilon})^{2n+1})$$

$$= \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k (a_k/k!) v_k(\bar{x}/\sqrt{\varepsilon}, t)$$

$$\sum_{k=0}^{2n} d_k H_k^*(\bar{x}/\sqrt{\varepsilon}, t) + O((\sqrt{\varepsilon})^{2n+1})$$

But $v_k(\bar{x}/\sqrt{\varepsilon}, t)$ is a solution to $u_t = \varepsilon u_{xx}$ and similar to (6) we have

$$v_k(\bar{x}/\sqrt{\varepsilon}, t) = \sum_{i=0}^{k'} \frac{(\varepsilon L)^i (\bar{x}/\sqrt{\varepsilon})^k t^i}{i!}$$

hence

$$u(x,t) = \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k (a_k/k!) \left(\sum_{i=0}^{k'} \frac{(\varepsilon L)^i (\bar{x}/\sqrt{\varepsilon})^k t^i}{i!} \right)$$

$$- \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k d_k H_k^*(\bar{x}/\sqrt{\varepsilon}, t) + 0((\sqrt{\varepsilon})^{2n+1})$$

Because $k \leq 2n$ implies $k' \leq n$, k' can be replaced by n. Then reversing the order of the summation gives

$$u(x,t) = \sum_{i=0}^{n} \left(\frac{1}{i!}\right) t^i (\varepsilon L)^i \left[\sum_{k=0}^{2n} (a_k/k!) \bar{x}^k \right]$$

$$- \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k d_k H_k^*(\bar{x}/\sqrt{\varepsilon}, t) + 0((\sqrt{\varepsilon})^{2n+1})$$

$$= \sum_{i=0}^{n} (1/i!) (\varepsilon t)^i L^i \left[\sum_{k=0}^{2n} (a_k/k!) \bar{x}^k \right]$$

$$- \sum_{k=0}^{2n} (\sqrt{\varepsilon})^k d_k H_k^*(\bar{x}/\sqrt{\varepsilon}, t) + O((\sqrt{\varepsilon})^{2n+1})$$

The expansion (13) of the initial value function can be used to simplify this for $x > x_o$

$$(16)\qquad u(x,t) = \sum_{i=0}^{n} \frac{(\varepsilon t)^i L^i(\phi(x))}{i!} - \sum_{k=0}^{2n} d_k (\sqrt{\varepsilon})^k H_k^*(\bar{x}/\sqrt{\varepsilon}, t)$$

$$+ O((\sqrt{\varepsilon})^{2n+1} \quad \text{where } \bar{x} = x - x_o$$

By similar derivation the corresponding formula for $x < x_o$ can be obtained:

$$(17)\qquad u(x,t) = \sum_{i=0}^{n} \frac{(\varepsilon t)^i L^i(\phi(x))}{i!} + \sum_{k=0}^{2n} d_k(\sqrt{\varepsilon})^k H_k(\bar{x}/\sqrt{\varepsilon}, t)$$

$$+ O((\sqrt{\varepsilon})^{2n+1})$$

In this form it is apparent how the inner solution given by (16) for $x > x_o$ and by (17) for $x < x_o$ is just the outer solution plus a correction for the interior layer. An analysis of the functions H_n and H_n^* would show that $H_n(x,t)$ becomes neglible faster than any power of $(1/x)$ as $x \to -\infty$ and similarly $H_n^*(x,t)$ becomes neglible as $x \to +\infty$. The correction terms in (16) and (17) are therefore neglible except in a neighborhood of the subcharacteristic $x = x_o$.

6. Solution for Piecewise Polynomial Initial Data. In previous sections asymptotic analysis was used to derive approximate solutions to (1), (2), (3) assuming that ε was small. In this section we show how an explicit solution in series form can be obtained for the same problem, regardless of the magnitude of ε, when $\phi(x)$ is a piecewise polynomial (with no continuity required between the pieces). Any initial value function can be approximated, as accurately as necessary by such functions and the maximum principle for parabolic equations guarantees that no greater error is thereby introduced into the solution. If εt is small the solution can be considerably simplified by dropping insignificant terms and in the limit we obtain the same asymptotic representations as before - (7), (16) and (17).

The solution will be obtained by writing the function $\phi(x)$ as a combination of the functions h_n and h_n^* and then expressing $u(x,t)$ as the same combination of H_n and H_n^*. We shall suppose there are points x_i such that $-\infty \leq \ldots < x_{-2} < x_{-1} < x_o < x_1 < \ldots \leq \infty$ and polynomials $p_i(x)$ with

$$\phi(x) = p_i(x) \qquad \text{in } [x_i, x_{i+1}).$$

The closed bracket is a convenience here since the value of ϕ at any one point is immaterial. The index i is assumed to vary over

some index set $I = J \cup K$ where $J = \{1,2,\ldots\}$ and $K = \{0,-1,-2,\ldots\}$. These sets may be finite or infinite but are non-void and we assume that only finitely many x_i are contained in any bounded interval.

For $k \geq 0$, let

$$n(k) = \max \{\text{degree } (p_i): i = 0, 1,\ldots,k\}$$

$$n(-k) = \max \{\text{degree } (p_i): i = 0, -1, -2,\ldots,-k\}$$

The representation of $\phi(x)$ has jumps at the points x_i so we define

$$d_i^n = \frac{1}{n!} [p_i^{(n)}(x_i) - p_{i-1}^{(n)}(x_i)] \qquad \text{for } x_i \neq \pm \infty$$

Let the coefficients of the polynomial $p_o(x)$ be denoted by a_i; that is

$$p_o(x) = \sum_{i=0}^{n(o)} a_i x^i$$

The initial values can now be expressed

$$\phi(x) = \sum_{i=o}^{n(o)} a_i x^i + \sum_{j\in J} \left(\sum_{n=0}^{n(j)} d_j^n h_n (x - x_j) \right) \tag{18}$$

$$- \sum_{k\in K} \left(\sum_{n=0}^{n(k)} d_k^n h_n^* (x - x_k) \right)$$

This series is well defined since for any given x only a finite number of terms are non-zero. (Recall $h_n (x - x_j) = 0$ if $x < x_j$ and $h_n^* (x - x_k) = 0$ if $x > x_k$.)

The functions $H_n(x - x_j, \varepsilon t)$ and $H_n^*(x - x_k, \varepsilon t)$ satisfy equation (1) and have the initial values $h_n(x - x_j)$ and $h_n^*(x - x_k)$ respectively. Using the linearity of the solutions of (1) we may therefore write a formal expression for $u(x,t)$

$$u(x,t) = \sum_{i=0}^{n(o)} \alpha_i v_i(x,\varepsilon t) + \sum_{j\in J} \sum_{n=0}^{n(j)} d_j^n H_n(x - x_j, \varepsilon t) \tag{19}$$

$$- \sum_{k\in K} \sum_{n=0}^{n(k)} d_k^n H_n^* (x - x_k, \varepsilon t)$$

When J and K are finite this representation is not merely formal but actually solves (1), (2), (3). When J or K is infinite the representation is still the real solution of the problem under certain general growth conditions on the polynomials $p_i(x)$. For this purpose we make the following definition:

The family of functions $\{p_i(x): i\in I = J \cup K\}$ is said to be outward bounded by $f(x)$ with respect to the points $\{x_i: i\in I=J\cup K\}$ if

$$|p_j(x)| \leq f(x) \quad \text{for all } x > x_j$$

$$|p_k(x)| \leq f(x) \quad \text{for all } x < x_{k+1}$$

Theorem: The formal series (19) converges for $0<t<1/4\varepsilon\alpha$ to the solution of (1), (2) (3) when $\phi(x)$ is given by (18) provided $\{p_i(x): i\in I = J\cup K\}$ is outward bounded by $2M \exp[\alpha x^2]$ with respect to the points $\{x_i: i\in I\}$ for some M, $\alpha \geq 0$.

Proof: Omitted. The argument is fairly straightforward using the integral representation of $u(x,t)$ and the properties of the fundamental solution $F(x,t)$.

7. Asymptotic Analysis. The solution (19) of problem (1), (2), (3) in general will contain an infinite number of terms. However most of these are neglibible even when εt is not small provided $x - x_j$ and $x - x_k$ are large enough.

This becomes apparent from studying the asymptotic behavior of H_n and H_n^* which can be summarized as follows:

a.) for fixed $x > 0$ $H_n^*(x,\varepsilon t)\to 0$ faster than any power of εt.

b.) for fixed $x < 0$ $H_n(x,\varepsilon t)\to 0$ faster than any power of εt.

c.) for $x > 0$ and $x \leq O(\sqrt{\varepsilon t})$ $H_n^*(x,\varepsilon t)$ is $O(\sqrt{\varepsilon t}^n)$.

d.) for $x < 0$ and $x \leq O(\sqrt{\varepsilon t})$ $H_n(x,\varepsilon t)$ is $O(\sqrt{\varepsilon t}^n)$.

In the representation (19) of $u(x,t)$ we assume without loss of generality that $x_0 \leq x < x_1$ (otherwise the indexes can be shifted). It follows that $x - x_j < 0$ for $j\in J$ and $x - x_k \geq 0$ for $k\in K$ with equality only possible for $k=0$. Using a.) - d.) we may conclude that terms $H_n(x - x_j, \varepsilon t)$ and $H_n^*(x - x_k, \varepsilon t)$ are neglible as $\varepsilon t\to 0$ unless $x - x_j$ and $x - x_k$ are $O(\sqrt{\varepsilon t})$. Therefore we define the sets J_ε and K_ε by

$$J_\varepsilon = \{j\in J: x - x_j = O(\sqrt{\varepsilon t})\}$$

$$K_\varepsilon = \{k\in K: x - x_k = O(\sqrt{\varepsilon t})\}$$

These sets must be finite since only finitely many x_j and x_k are contained in any bounded interval. For asymptotic approximations to $U(k,t)$ as $\varepsilon \to 0$ we can restrict the index sets J and K to J_ε and K_ε. Furthermore, for N-th order accuracy the terms H_n and H_n^* can be neglected if $n > N$, because of c.) and d.). The representation (19) can therefore be written as

$$(20)\qquad u(x,t) = \sum_{i=0}^{N} \alpha_i\, v_i(x,\varepsilon t) + \sum_{j\in J_\varepsilon} \sum_{n=0}^{N} d_j^n\, H_n(x-x_j,\varepsilon t)$$

$$- \sum_{k\in K_\varepsilon} \sum_{n=0}^{N} d_k^n\, H_n^*(x-x_k,\varepsilon t) + R_N$$

where the remainder $R_N = O(\sqrt{\varepsilon t})^{N+1}$ as $\varepsilon t \to 0$.

The sets J_ε and K_ε are monotone decreasing with respect to ε (in the set theoretic sense) and eventually reduce to at most one element. The resulting form of (20) is then equivalent to the inner solution representations (16), for $x - x_1 = O(\sqrt{\varepsilon t})$, or (17), for $x - x_o = O(\sqrt{\varepsilon t})$ and (7) otherwise.

8. Mixed Boundary - Initial Value Problems. One of the techniques for solving mixed boundary - initial value problems is to convert the problem into an appropriate pure initial value problem for which the fundamental solution is known. The solution of the new problem is then shown to satisfy all of conditions of the original problem. This approach is particularly useful in the present case since we have developed all of the necessary tools for solving the Cauchy problem.

Unfortunately there does not seem to be any standard reference for this technique and we do not have space here for discussing the matter. Therefore we shall only indicate the proper initial value problem which can be used to solve certain mixed boundary - initial value problems. Greater detail will be included in [5] and is being submitted for publication in a separate paper.

The following list is valid for the equation $u_t = \varepsilon u_{xx}$.

A. Semi-infinite domain, zero boundary condition

Mixed BVP-IVP:	$u(x,0) = \phi(x)$	$x > 0$
	$u(o,t) = 0$	$t > 0$
Corresponding IVP:	$u(x,o) = \phi(x)$	$x > 0$
	$u(x,o) = -\phi(-x)$	$x < 0$

B. Semi-infinite domain, zero initial condition

Mixed BVP-IVP:	$u(x,o) = 0$	$x > 0$
	$u(o,t) = h_n(t)$	$t > 0$
Corresponding IVP:	$u(x,o) = 2h_{2n}^*(x/\sqrt{\varepsilon})$	

C. Finite domain, zero boundary conditions

Mixed BVP-IVP:
$$\begin{aligned} u(x,o) &= h_n(x) && 0 < x < x_o \\ u(o,t) &= 0 && t > 0 \\ u(x_o,t) &= 0 && t > 0 \end{aligned}$$

Corresponding IVP:
$$u(x,o) = h_n(x-2k\,x_o) - h_n^*(x-2k\,x_o)$$
$$\text{for } (2k-1)\,x_o < x < (2k+1)x_o$$

D. Finite domain, one non-zero boundary condition

Mixed BVP-IVP:
$$\begin{aligned} u(x,o) &= o && 0 < x < x_o \\ u(o,t) &= h_n(t) && t > 0 \\ u(x_o,t) &= 0 && t > 0 \end{aligned}$$

Corresponding IVP:
$$u(x,o) = 2 \sum_{j=0}^{\infty} h_{2n}^* \left(\frac{x+2jx_o}{\sqrt{\varepsilon}}\right) - 2 \sum_{k=0}^{\infty} h_{2n} \left(\frac{x-2k\,x_o}{\sqrt{\varepsilon}}\right)$$

These correspondences combined with the theory developed in previous sections for the Cauchy problem allow us to solve the heat conduction problem in a finite rod explicitly when the data is given in piecewise polynomial form.

8. Conclusions. Basically what we have developed in this discussion is an alternative to the Fourier method for solving the diffusion equation. The present method, based on asymptotic analysis, is accurate for small εt and is particularly advantageous when any sort of discontinuities are present in the initial and boundary data. In general very few terms are needed to give accurate estimates of the solution. To contrast this with the Fourier approach one should note how many terms are required to resolve a discontinuous function into sines and cosines with reasonable accuracy. Although the higher frequency modes tend to cancel out and become insignificant with time this is not helpful in describing the early development of the solution.

In our opinion the present method is a far more natural approach to diffusion problems when steep gradients are present. It also has the advantage of providing formulae rather than numbers for describing the physical behavior.

REFERENCES

[1] Kamenomostskaya, S.L., "On Equations of Elliptic and Parabolic Type with a Small Parameter in the Highest Derivatives". Mat. Sbornik, N.S., 31, (73), 703-708 (1952).

[2] Aronson, D.G., "Linear Parabolic Equations Containing a Small Parameter". J. Rat. Mech. and Anal. 5, 1003-1014 (1956).

[3] Bobisud, L.E., "Second-Order Linear Parabolic Equations with a Small Parameter". Archiv Rat. Mech. Anal. 27, 385-397 (1968).

[4] Bobisud, L.E., "Parabolic Equations with a Small Parameter and Discontinuous Data". J. Math. Anal. Appl. 26, 208-220 (1969).

[5] Polk, J.F., "Singular Perturbation Analysis in Diffusion Equations", Ph.D Dissertation, U. of Delaware, not yet completed.

[6] Friedman, A., "Partial Differential Equations of Parabolic Type". Prentice-Hall, Englewood Cliffs 1964.

[7] Rosenbloom, P.C. and Widder, D.V., "Expansions in Terms of Heat Polynomials and Related Functions". Trans. AMS(92), 220-266 (1959).

SAMPLE SIZES FOR MISSILE IN FLIGHT RELIABILITY DETERMINATION

Edward F. Southworth
Army Missile Test and Evaluation Directorate (ARMTE)
White Sands Missile Range, New Mexico 88002

ABSTRACT. This paper presents an unusual approach to determining sample sizes required to adequately measure reliability. This approach has been tailored to the specific problem of measuring the in flight reliability of guided missiles. It should also have application to other types of problems. The paper includes the rationale used to develop a measure of the probability of a successful reliability test. It then derives equations for calculating the probability of a successful reliability test. Finally the paper presents plots of the probability of a successful reliability test as a function of sample size for a wide range of reliability testing goals.

1. INTRODUCTION. As the name suggests, the primary function of the Army Missile Test and Evaluation Directorate (ARMTE) is test and evaluation of Army guided missile systems. A small, but important, part of this function is the determination of the reliability of these systems. Of particular concern is the determination of the in-flight reliability of the missiles themselves. There are three reasons for this concern.

a. First, the unit cost of most guided missiles is very high. In addition, missiles are produced in large quantity. Many more missiles are produced than other portions of most missile systems for the same reason that many more rounds of ammunition are produced than the guns that fire them. Initiating large scale production of a specific missile configuration requires a major decision. Naturally, the decision makers want assurnace that missile performance, including reliability, will be satisfactory before committing large sums to missile production.

b. Second, most modern missiles are configured as certified rounds. The missiles are sealed at the factory. They are not designed to be maintained in the field. The field army is neither trained, equipped, nor authorized to maintain them. Any problems discovered after they are manufactured can only be corrected by very expensive retrofit programs.

c. Third, there is a lot more than money involved. Mistaken estimates of missile performance contribute to mistakes in the tactical doctrine developed for optimum use of our weapon systems. Large discrepancies in our estimate of in flight reliability could lead to serious miscalculation of our defense posture. Such miscalculations can be the cause of tactical blunders.

Since measuring missile in flight reliability is so important, why not just measure it? The answer, of course, is that missile in flight reliability can only be measured in very expensive missile firing tests. These tests are costly because of the high cost of the missiles which must be expended, because of the high cost of the range support required, and because of the high cost of the targets needed for many of these tests.

It is obvious that an adequate measure of missile in flight reliability must be obtained with a minimum number of firings. The situation justifies more complex test planning techniques than are usually employed in designing reliability experiments.

There is one other unusual feature about this problem. ARMTE is responsible for measuring missile in flight reliability, but ARMTE is not responsible for making any decisions on the adequacy of this reliability. ARMTE is not involved in determining acceptance or rejection criteria, buyers' risks, or sellers' risks. For this reason, the standard operating characteristics curves have limited value.

This paper will first discuss ARMTE's reliability testing goals. It will then define a measure of the effectiveness of test designs in achieving these reliability testing goals. Finally, it will present families of graphs which plot the probability of a successful reliability test as a function of sample size for a wide variety of reliability testing goals.

2. RELIABILITY TESTING GOALS. The Army Missile Test and Evaluation Directorate has three types of missile in flight reliability testing goals. The first goal is to demonstrate that reliability is satisfactory. If we can accomplish that, all parties are happy and the Army obtains reliable missiles. To do this we must produce convincing evidence that reliability equals or exceeds whatever level is accepted as satisfactory.

The second goal is the opposite of the first. If we are unable to demonstrate that reliability is satisfactory, the possibility that reliability falls short of the satisfactory level must be investigated.

The third goal is addressed if we cannot accomplish either of the first two goals. If we cannot demonstrate that reliability is satisfactory, and we also fail to demonstrate that reliability is not satisfactory, then our goal is to demonstrate that reliability is near the satisfactory level. Our ability to accomplish this goal is dependent upon the effectiveness of the test design used to accomplish the first two goals. If the test design provides a high probability of detecting small differences between actual reliability and the satisfactory level, it is also capable of demonstrating that reliability is very near the satisfactory level. In such a test, failure to demonstrate that reliability is either satisfactory or unsatisfactory provides high confidence that it is near the satisfactory level. On the other hand, test designs which can detect

only large variations from the satisfactory reliability level can only define the value of reliability within broad bands of uncertainty.

3. A MEASURE OF TEST DESIGN EFFECTIVENESS. Now that ARMTE's reliability testing goals have been described, a measure of the effectiveness of test designs in achieving these goals will be defined. A test which achieves ARMTE's reliability testing goals must satisfy two criteria. First, it must produce a useful conclusion about reliability. The conclusion which the test produces can satisfy any one of the three reliability testing goals described in paragraph 2, above, to be useful. Second, the conclusion the test produces must be a correct conclusion. Based upon these criteria, the probability of a successful ARMTE reliability test is defined as the probability that the test produces a conclusion which satisfies one of ARMTE's reliability testing goals, multiplied by the conditional probability that the conclusion is correct.

4. DETERMINING THAT THE CONCLUSION IS CORRECT. A major part in predicting the probability of a successful ARMTE reliability test is determining the probability that a conclusion about reliability is correct. A good measure of the correctness of a conclusion about reliability is the statistical confidence which can be placed in that conclusion. Given the results of a reliability test in which there were s successes in n trials, it is easy to calculate the statistical confidence that actual reliability is greater than or equal to the satisfactory level. It is also easy to calculate the statistical confidence that actual reliability is less than the satisfactory level. Statistical confidence is calculated using the binomial distribution. The equations for calculating statistical confidence using the binomial distribution are well known. However, ready reference to these equations is needed for the further development of this paper. Therefore, these equations, and their rationale will be briefly reviewed.

Figure 1 shows the equations used in calculating confidence and provides some examples which explain the significance of the confidence values calculated.

Equation 1 is merely a statement of the most basic equation of the binomial distribution. It states that the probability of exactly "i" successes in "n" trials, if the probability of success in each trail is "p", equals n factorial, times p to the i^{th} power, times (1 - p) to the $n - i^{th}$ power, divided by i factorial, and also divided by (n - i) factorial. 0 factorial is defined as equal to 1.

The second column under examples shows the probability of achieving various numbers of successes in 25 trials if the probability of success in each trial equals .85. For example the probability of achieving exactly 21 successes under these conditions equals .2110.

FIGURE 1

EQUATIONS:

1. $P(i, n, p) = \frac{n!\, p^{i}(1 - p)^{n - i}}{i!\,(n - i)!}$ where $0! = 1$

2. $\sum_{i = s}^{n} P(i, n, p) = \sum_{i = s}^{n} \frac{n!\, p^{i}(1 - p)^{n - i}}{i!\,(n - i)!}$

3. $C(r \geq p_c) = 1 - \sum_{i = s}^{n} P(i, n, p_c) = 1 - \sum_{i = s}^{n} \frac{n!\, p_c^{\,i}(1 - p_c)^{n - i}}{i!\,(n - i)!}$

4. $C(r < p_c) = \sum_{i = s + 1}^{n} P(i, n, p_c) = \sum_{i = s + 1}^{n} \frac{n!\, p_c^{\,i}(1 - p_c)^{n - i}}{i!\,(n - i)!}$

EXAMPLES:

Number Of Successes in 25 Trials	Probability If p=.85 $P(i,n,p)=\frac{n!\,p^{i}(1-p)^{n-i}}{i!\,(n-i)!}$	Cumulative Probability If p = .85 $\sum_{i=s}^{n} P(i,n,p)=\sum_{i=s}^{n}\frac{n!\,p^{i}(1-p)^{n-i}}{i!\,(n-i)!}$	Confidence That R≥.85 $C(R\geq p_c)=1-\sum_{i=s}^{n} P(i,n,p_c)$	Confidence That R<.85 $C(R<p_c)=\sum_{i=s+1}^{n} P(i,n,p_c)$	Point Estimate Of p
25	.0172	.0172	.9828	.0000	1.00
24	.0759	.0931	.9069	.0172	.96
23	.1607	.2537	.7463	.0931	.92
22	.2174	.4711	.5289	.2537	.88
21	.2110	.6821	.3179	.4711	.84
20	.1564	.8385	.1615	.6821	.80
19	.0920	.9305	.0695	.8385	.76
18	.0441	.9745	.0255	.9305	.72
17	.0175	.9920	.0080	.9745	.68
16	.0058	.9979	.0021	.9920	.64
15	.0016	.9995	.0005	.9979	.60
14	.0004	.9999	.0001	.9995	.56
13	.0001	1.0000	.0000	.9999	.52
12	.0000	1.0000	.0000	1.0000	.48

Equation 2 is basic also. It is a statement of the cumulative probability from the binomial distribution. It shows that the probability of s or more successes in n trials, if the probability of successes in each trial equals p, is equal to the sum of the probability of s successes in n trials, plus the probability of s + 1 successes in n trials, and so on until the probability of n successes in n trials has been included.

The third column under examples shows the cumulative probability of s or more successes in 25 trials, if the probability of success in each trial equals .85, for various values of s. Each cumulative probability in column 3 equals the sum of the individual probabilities in column 2 associated with that number or more successes. For example, the probability of 21 or more successes under these conditions is .6821. This is the sum of the probability of exactly 21 successes, plus the probability of exactly 22 successes, and so on until the probability of 25 successes in 25 trials has been added.

Equation 3 requires a little explanation. It is used to calculate confidence that reliability equals or exceeds some criterion. In this case, the criterion is the satisfactory level for in flight reliability. The far left term in the equation "C $(r \geq (p_c)$" means "confidence that reliability equals or exceeds criterion". The symbol "p_c" designates the probability of success in each trial which satisfies the criterion. It is interesting to note that the center and right terms of equation 3 are the complements of the corresponding terms of equation 2. Equation 2 defines the probability that s or more successes would occur in n random trials if the probability of success in each trial equals "p". Equation 3 is the exact opposite. It defines the confidence we can establish that a given result did not occur from random sampling. This confidence, as shown, equals unity minus the probability that the result would occur from random sampling. Equation 3 is used to calculate confidence that a given test result demonstrates that reliability equals or exceeds a satisfactory level.

Column 4, under examples, shows the confidence that reliability equals or exceeds .85 if we experience s successes in 25 trials. As you would expect, each value in column 4 equals unity minus its corresponding value in column 3.

Equation 4 is used to calculate confidence that reliability is less than the criterion. It is very similar to equation 2. The only difference is that "i" is summed from s + 1 to n in equation 4 whereas "i" is summed from "s" to n in equation 2. Equation 2 is used to determine the probability of s or more successes in n trials if the probability of success in each trial equals p. The principal of equation 4 is that this same probability is the confidence that reliability is less than p if we experience less than s successes in n trials. This can most easily be illustrated by example. Column 3 under examples shows that the probability of 21 or more successes in 25 trials if p equals .85 is .6821. Therefore if we

experience fewer than 21 successes in 25 trials we can have a confidence of .6821 that p is less than .85. Column 5 under examples shows this confidence value for the case in which we experience only 20 successes in 25 trials. In fact every entry in column 5 is numerically equal to the next higher entry in column 3.

There is one more point that I would like to discuss before leaving Figure 1. The far right column under examples shows the point estimate for reliability for each number of successes in 25 trials. As you can see, the point estimate is merely the numerical equivalent of the number of successes divided by the number of trials. It is important to note that whenever the point estimate exceeds the criterion for reliability then confidence that reliability exceeds criterion is greater than confidence that reliability is below the criterion. In the examples, the criterion is .85. When 22 successes are obtained in 25 trials the point estimate is .88. In this case confidence that reliability exceeds .85 is shown as .5289. Confidence that reliability is below .85 equals only .2537. On the other hand, if there are only 21 successes in 25 trials the point estimate is .84 which is below the criterion of .85. Note that when the point estimate is below the criterion there is greater confidence that reliability is below the criterion than the confidence that it is above. This example illustrates a rule that applies to all cases.

5. PROBABILITY OF A SUCCESSFUL TEST: The first four equations on Figure 2 are merely repetitions of the four equations just discussed on Figure 1. They are repeated to lend continuity to this discussion. They provide an important link from the familiar to the unfamiliar represented by equation 5. Equation 5 is the first unusual feature to the approach presented in this paper.

The left side of equation 5, "Pst (n, r, p_c)", means "probability of a successful test" if the sample size equals "n", the actual reliability equals "r" and the probability of success in each trial which satisfies the criterion equals "p_c". The right side of the equation is a double summation. It's similar to a double integral except that it applies to sample sizes which can only be whole numbers rather than continuous variables. Equation 5 is the method used to predict the probability of a reliability test which successfully accomplishes our first goal of demonstrating that actual reliability equals or exceeds the satisfactory level. In other words, it predicts the probability that we can successfully demonstrate that "r" equals or exceeds "p_c".

The right side of equation 5 is divided into two major parts. The part which immediately follows the first summation sign calculates the probability of achieving exactly s successes in n trials if the probability of success in each trial (i. e. reliability) equals "r". You can see that this part of the equation is in the same form as equation 1 which is used for the same type of calculation. The last part of equation 5

calculates confidence that reliability equals or exceeds the criterion, "p_c", if we experience exactly "s" successes in "n" trials. You can see that this part of the equation is in the same form as equation 3 which is used for the same type of calculation. The two parts, together, equal the probability of achieving "s" successes in "n" trials, multiplied by the confidence provided by "s" successes in "n" trials that reliability equals or exceeds the criterion, "p_c".

The remaining portion of equation 5 is the first summation sign. This summation sign shows that the calculation described above is repeated for different values of "s" and that the results are added to obtain the probability of a successful test. All values of "s" are used which equal or exceed the sample size multipled by the criterion, "p_c". Earlier I pointed out that all values of "s" which equal or exceed this quantity contribute to our confidence that reliability is satisfactory.

FIGURE 2

EQUATIONS

1. $$P(i, n, p) = \frac{n!\, p^i (1 - p)^{n - i}}{i!\,(n - i)!} \quad \text{where } 0! = 1$$

2. $$\sum_{i = s}^{n} P(i, n, p) = \sum_{i = s}^{n} \frac{n!\, p^i (1 - p)^{n - i}}{i!\,(n - i)!}$$

3. $$C(r \geq p_c) = 1 - \sum_{i = s}^{n} P(i, n, p_c) = 1 - \sum_{i = s}^{n} \frac{n!\, p_c^{\,i} (1 - p_c)^{n - i}}{i!\,(n - i)!}$$

4. $$C(r < p_c) = \sum_{i = s + 1}^{n} P(i, n, p_c) = \sum_{i = s + 1}^{n} \frac{n!\, p_c^{\,i} (1 - p_c)^{n - i}}{i!\,(n - i)!}$$

5. $$P_{ST}(n, r, p_c) = \sum_{s \geq np_c}^{n} \left[\frac{n!\, r^s (1 - r)^{n - s}}{s!\,(n - s)!} \left(1 - \sum_{i=s}^{n} \frac{n!\, p_c^{\,i} (1 - p_c)^{n-i}}{i!\,(n - i)!} \right) \right]$$

Figure 3 shows one of the consequences of this relationship. This Figure lists the values of "s" which contribute to our confidence that reliability is satisfactory for various numbers of samples and values of criteria. It is evident that there are more values of "s" which contribute as sample size is increased and as the criterion "p_c", is reduced. Each condition under which an additional value of "s" is introduced into the calculation is marked by a pip on Figure 3. Each time an additional value of "s" is introduced, there is a discontinuity in the graphs which show the probability of a successful test. The result is that these graphs have a sawtooth characteristic to them. The next figure will demonstrate this characteristic.

Figure 4 shows a family of graphs which plot the probability of a successful test as a function of sample size. Each graph is applicable to the goal of demonstrating that reliability is equal to or greater than .95. The bottom graph is a plot of the probability of achieving this goal if the actual probability of success in each trial (P_a) equals .95. In other words, actual reliability equals the criterion. The other graphs apply to cases in which actual reliability exceeds the criterion. Graphs are provided for reliability values of .96, .97, .98 and .99.

There are several features of these graphs which require explanation. First, they are plotted as though they are continuous functions. This was done to connect together successive points associated with the same actual probability of success in each trial. Obviously, the only points on the graphs that have any meaning are those associated with whole numbers corresponding to discrete sample sizes. Second, the sawtooth shape of these graphs has already been explained, but further explanation would probably be helpful now that you can see this characteristic. The phenomena, loosely referred to as discontinuities, occur at sample sizes which are multiples of 20 if the goal is to demonstrate that reliability is equal to or greater than .95. The reason is that the probability of achieving 19 or more successes in 20 trials is far greater than the probability of achieving 19 successes in 19 trials. Similarly, the probability of achieving 38 or more successes in 40 trials is greater than the probability of achieving 38 or more successes in 39 trials. At every multiple of 20 samples the experiment can accomodate one additional failure without driving the point estimate below .95. Hence the sawtoothed shape of these graphs. There is a third feature of these graphs that requires explanation. The graphs show that there are conditions in which the probability of a successful test decreases slightly as sample size increases slightly. This phenomenon is very disturbing to most people. It defies intuition. However, it actually happens. If the actual reliability is greater than the satisfactory level, the long term trend is always to increase the probability of a successful test with large increases in sample size. However, the sawtoothed shape of these graphs causes small intervals where the graphs have negative slopes when the actual reliability is only slightly greater than the satisfactory level.

FIGURE 3

	p_c = .95		p_c = .85		p_c = .75	
n	n p_c	Values of "S" $\geq$ n p_c	n p_c	Values of "S" $\geq$ n p_c	n p_c	Values of "S" $\geq$ n p_c
1	.95	1	.85	1	.75	1
2	1.90	2	1.70	2	1.50	2
3	2.85	3	2.55	3	2.25	3
4	3.80	4	3.40	4	3.00	4,3
5	4.75	5	4.25	5	3.75	5,4
6	5.70	6	5.10	6	4.50	6,5
7	6.65	7	5.95	7,6	5.25	7,6
8	7.60	8	6.80	8,7	6.00	8,7,6
9	8.55	9	7.65	9,8	6.75	9,8,7
10	9.50	10	8.50	10,9	7.50	10,9,8
11	10.45	11	9.35	11,10	8.25	11,10,9
12	11.40	12	10.20	12,11	9.00	12,11,10,9
13	12.35	13	11.05	13,12	9.75	13,12,11,10
14	13.30	14	11.90	14,13,12	10.50	14,13,12,11
15	14.25	15	12.75	15,14,13	11.25	15,14,13,12
16	15.20	16	13.60	16,15,14	12.00	16,15,14,13,12
17	16.15	17	14.45	17,16,15	12.75	17,16,15,14,13
18	17.10	18	15.30	18,17,16	13.50	18,17,16,15,14
19	18.05	19	16.15	19,18,17	14.25	19,18,17,16,15
20	19.00	20,19	17.00	20,19,18,17	15.00	20,19,18,17,16,15
21	19.95	21,20	17.85	21,20,19,18	15.75	21,20,19,18,17,16
22	20.90	22,21	18.70	22,21,20,19	16.50	22,21,20,19,18,17
23	21.85	23,22	19.55	23,22,21,20	17.25	23,22,21,20,19,18
24	22.80	24,23	20.40	24,23,22,21	18.00	24,23,22,21,20,19,18
25	23.75	25,24	21.25	25,24,23,22	18.75	25,24,23,22,21,20,19

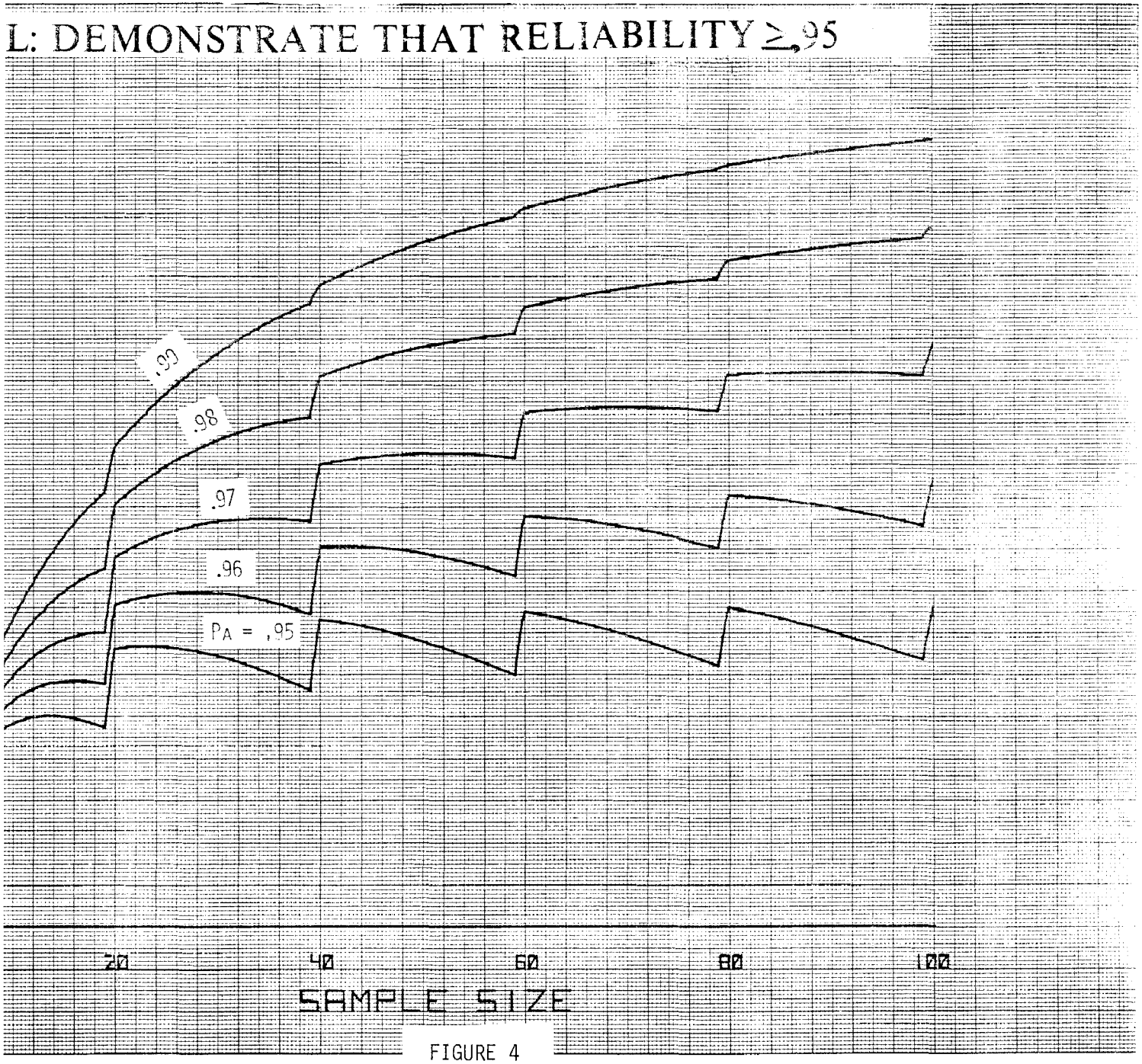

L: DEMONSTRATE THAT RELIABILITY ≥ .95
.99
.98
.97
.96
PA = .95
20
40
60
80
100
SAMPLE SIZE

FIGURE 4

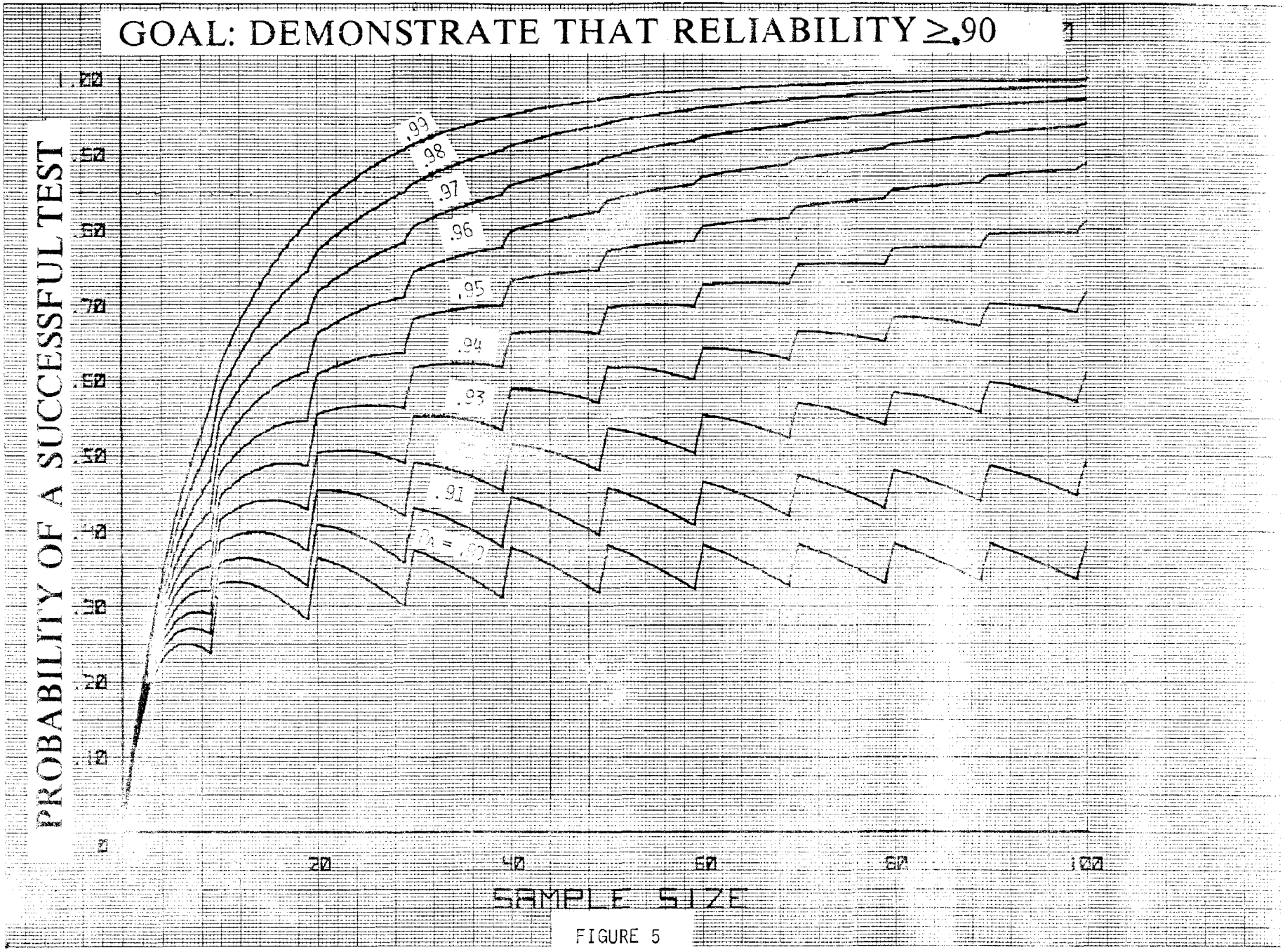

GOAL: DEMONSTRATE THAT RELIABILITY ≥.90
PROBABILITY OF A SUCCESSFUL TEST
.99
.98
.97
.96
.95
.94
.93
.91
20
40
60
80
100
SAMPLE SIZE

FIGURE 5

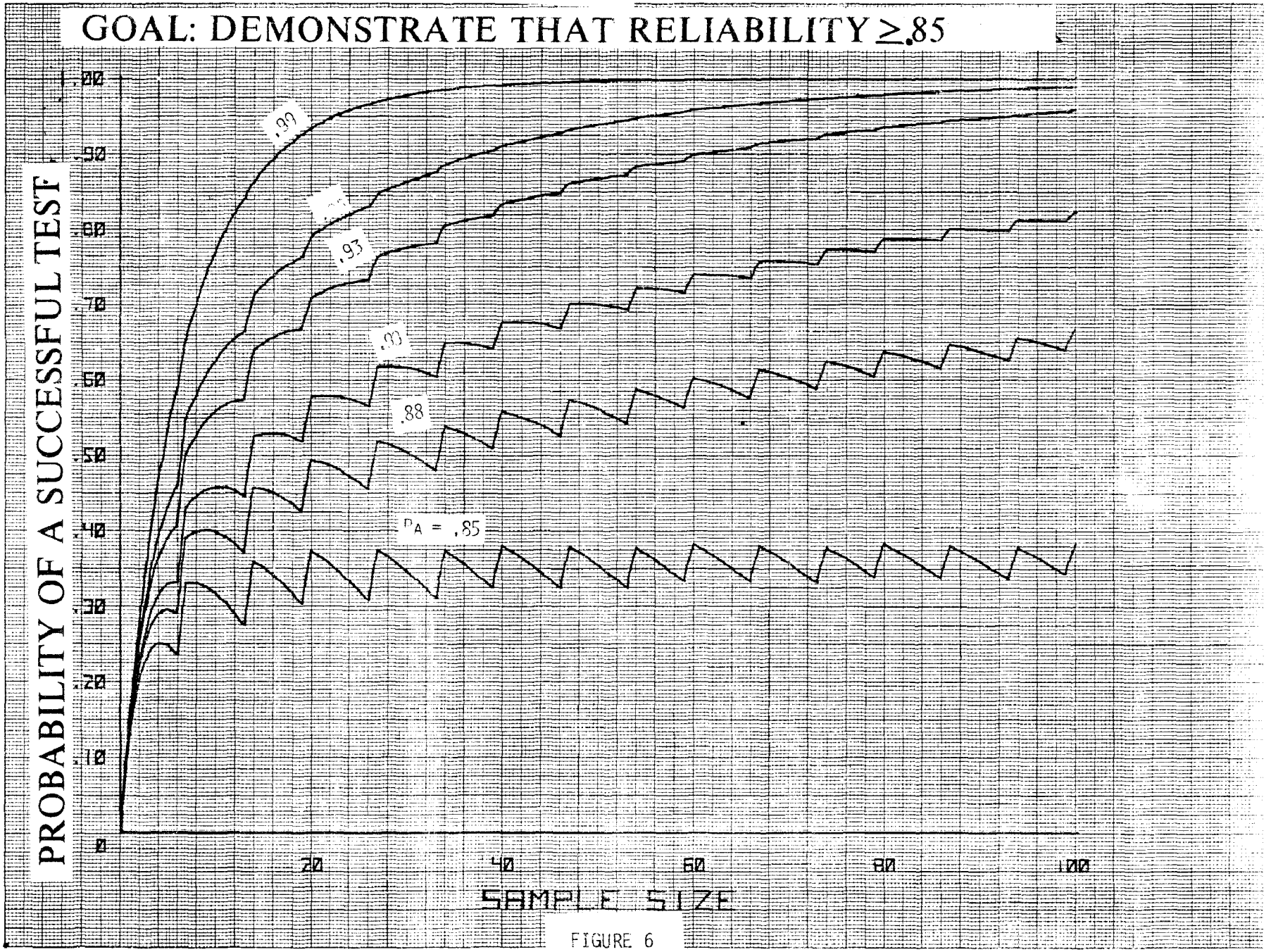

GOAL: DEMONSTRATE THAT RELIABILITY ≥.85
PROBABILITY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.99
.93
.88
P_A = .85
20
40
60
80
100
SAMPLE SIZE

FIGURE 6

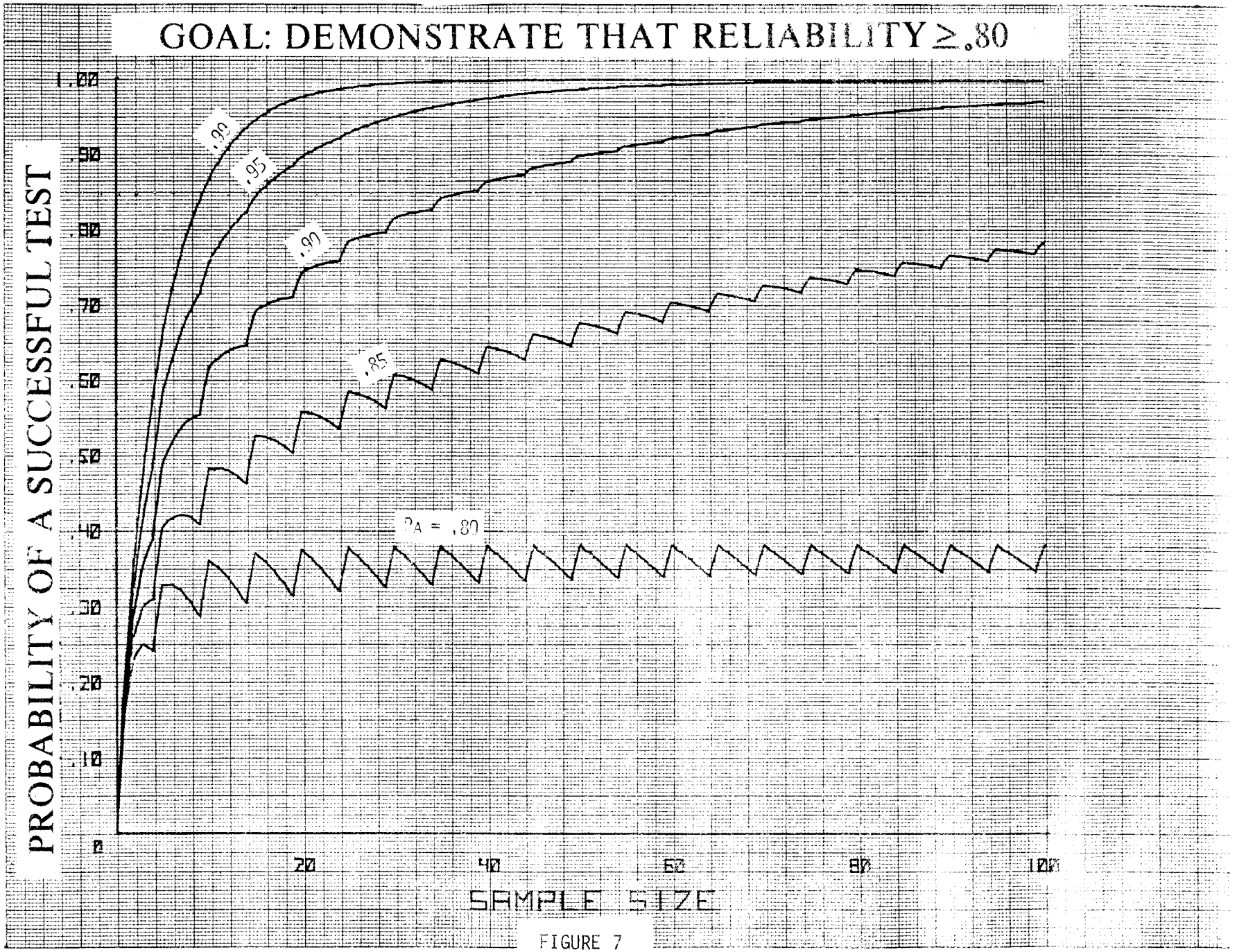

GOAL: DEMONSTRATE THAT RELIABILITY ≥ .80
PROBABILITY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.99
.95
.90
.85
P_A = .80
20
40
60
80
100
SAMPLE SIZE

FIGURE 7

GOAL: DEMONSTRATE THAT RELIABILITY ≥.75
PROBABILTY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.99
.95
.90
.85
.80
PA = .75
20
40
60
80
SAMPLE SIZE

FIGURE 8

GOAL: DEMONSTRATE THAT RELIABILITY ≥ .70
PROBABILITY OF A SUCCESSFUL TEST
SAMPLE SIZE
P_A = .70
.75
.80
.85
.90
.95
.99
1.00
.90
.80
.70
.60
.50
0
20
40
60
80
100

FIGURE 9

GOAL: DEMONSTRATE THAT RELIABILITY ≥.65
PROBABILTY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.99
.95
.90
.85
.80
.75
.70
PA = .65
20
40
60
80
100
SAMPLE SIZE

FIGURE 10

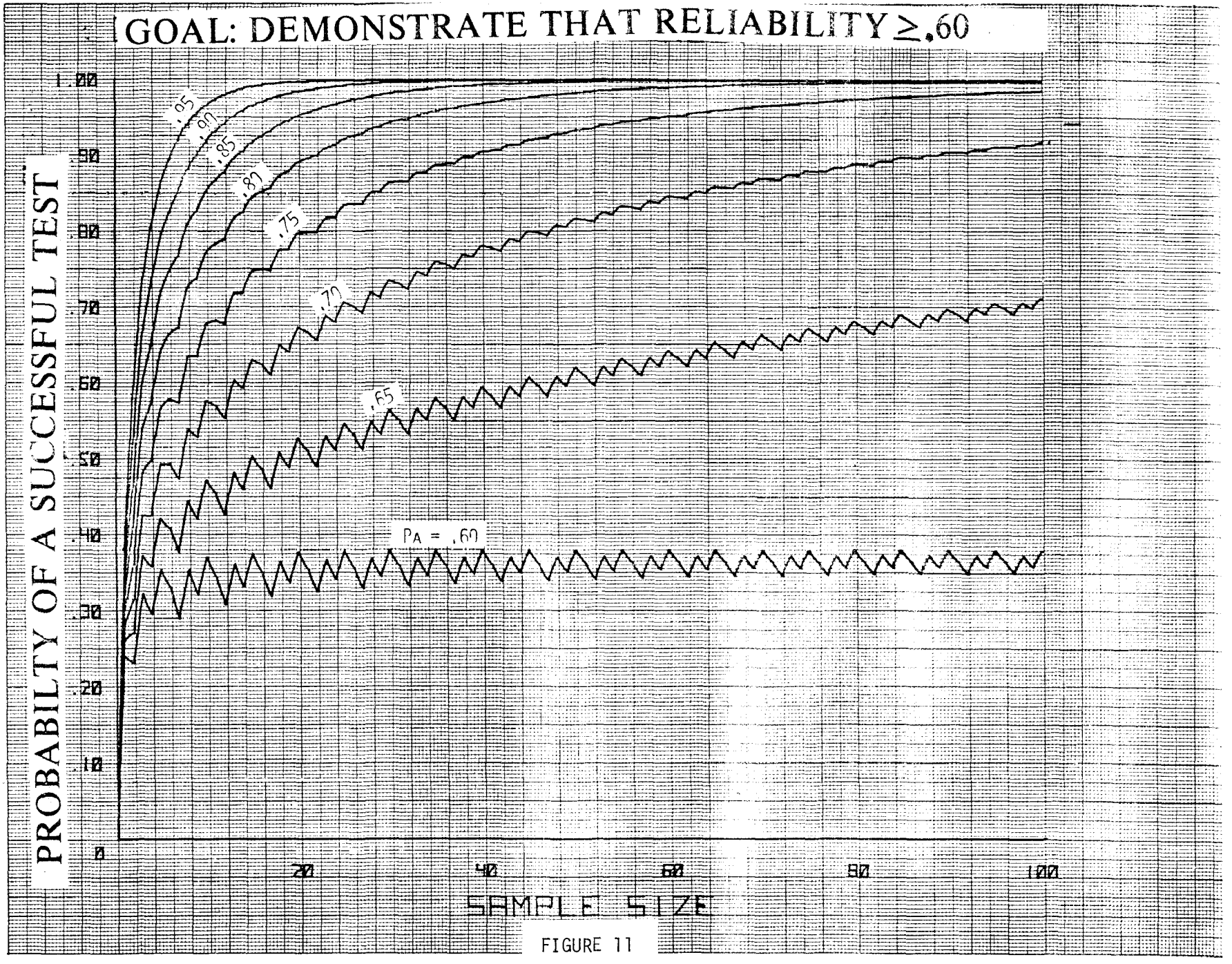

GOAL: DEMONSTRATE THAT RELIABILITY ≥ .60
PROBABILTY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.95
.90
.85
.80
.75
.70
.65
PA = .60
20
40
60
80
100
SAMPLE SIZE

FIGURE 11

Figures 5 thru 11 are similar to Figure 4 except that they apply to different levels of satisfactory reliability.

They cover goals of demonstrating that reliability is equal to or better than .90, .85, .80, .75, .70, .65, and .60.

Figure 12 adds a sixth and final equation to the group. Equation 6 is the counterpart to equation 5. Equation 5 is used to determine the sample size required to achieve our first goal of demonstrating that reliability is satisfactory. Equation 6 is used to determine the sample size required to achieve the second goal of demonstrating that reliability is unsatisfactory. The right side of equation 6 is divided into two major parts, as equation 5 was. The part which immediately follows the first summation sign calculates the probability of achieving exactly s successes in n trials if the probability of success in each trial (i. e. reliability) equals r. You can see that this part of the equation is in the same form as the corresponding part of equation 5 and equation 1, which are used for the same type of calculation. The last part of equation 6 calculates confidence that reliability is below the criterion, "p_c", if we experience exactly s successes in n trials. You can see that this part of the equation is in the same form as equation 4, which is used for the same type of calculation. The two parts, together, equal the probability of achieving s successes in n trials, multiplied by the confidence this result provides that reliability is below the criterion, p_c.

The first summation sign in equation 6 shows that the calculation just described is summed from s = 0 to the highest value of s which satisfies the relationship that s is equal to or less than the number of trials multiplied by the satisfactory level of reliability. In other words, all values of s are summed for which the point estimate of reliability is below the criterion. Thus, all values of s are summed which contribute confidence that reliability is unsatisfactory.

Figure 13 is similar to some of the previous figures except that this one shows the probability of demonstrating that reliability is less than the criterion of .95. As you can see, the graphs on this Figure also have sawtoothed shapes. They also peak out at sample sizes which are multiples of 20 and then drop when the sample size is increased by one. On these graphs, also, the prevailing trend is for the probability of a successful reliability test to increase as sample size increases. Again, though, there are some exceptions.

Figures 14 through 20 are similar to this Figure except that they apply to different levels of satisfactory reliability.

They cover satisfactory reliability levels of .90, .85, .80, .75, .70, .65, and .60.

5. CONCLUSIONS.

FIGURE 12

EQUATIONS

1. $$P(i, n, p) = \frac{n!\, p^{i}\, (1 - p)^{n - i}}{i!\, (n - i)!} \qquad \text{where } 0! = 1$$

2. $$\sum_{i = s}^{n} P(i, n, p) = \sum_{i = s}^{n} \frac{n!\, p^{i}\, (1 - p)^{n - i}}{i!\, (n - i)!}$$

3. $$C(r \geq p_c) = 1 - \sum_{i = s}^{n} P(i, n, p_c) = 1 - \sum_{i = s}^{n} \frac{n!\, p_c^{\,i}\, (1 - p_c)^{n - i}}{i!\, (n - i)!}$$

4. $$C(r < p_c) = \sum_{i = s + 1}^{n} P(i, n, p_c) = \sum_{i = s + 1}^{n} \frac{n!\, p_c^{\,i}\, (1 - p_c)^{n - i}}{i!\, (n - i)!}$$

5. $$P_{ST}(n, r, p_c) = \sum_{s \geq np_c}^{n} \left[\frac{n!\, r^{s}\, (1 - r)^{n - s}}{s!\, (n - s)!} \left(1 - \sum_{i = s}^{n} \frac{n!\, p_c^{\,i}\, (1 - p_c)^{n - i}}{i!\, (n - i)!} \right) \right]$$

6. $$P_{ST}(n, r, p_c) = \sum_{s = 0}^{s \leq np_c} \left(\frac{n!\, r^{s}\, (1 - r)^{n - s}}{s!\, (n - s)!} \sum_{i = s + 1}^{n} \frac{n!\, p_c^{\,i}\, (1 - p_c)^{n - i}}{i!\, (n - i)!} \right)$$

a. It is practical to plot the probabilities of achieving a variety of reliability testing goals as a direct function of sample size.

b. Such plots are of value to the test designer because they provide him a direct view of the trade off between his goals and their cost in additional samples.

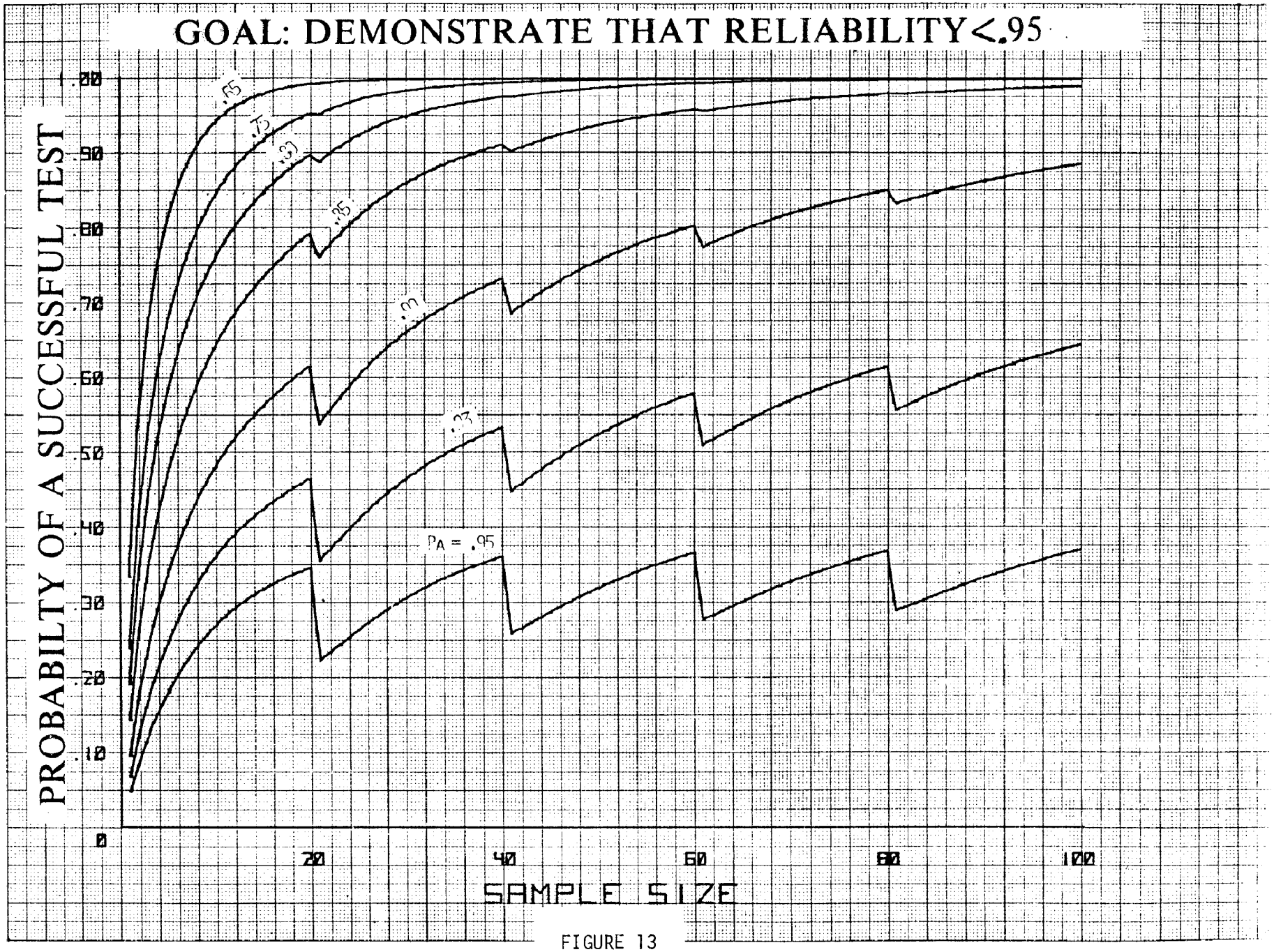

GOAL: DEMONSTRATE THAT RELIABILITY<.95
PROBABILTY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.95
.75
.33
.25
.13
.33
P_A = .95
20
40
60
80
100
SAMPLE SIZE

FIGURE 13

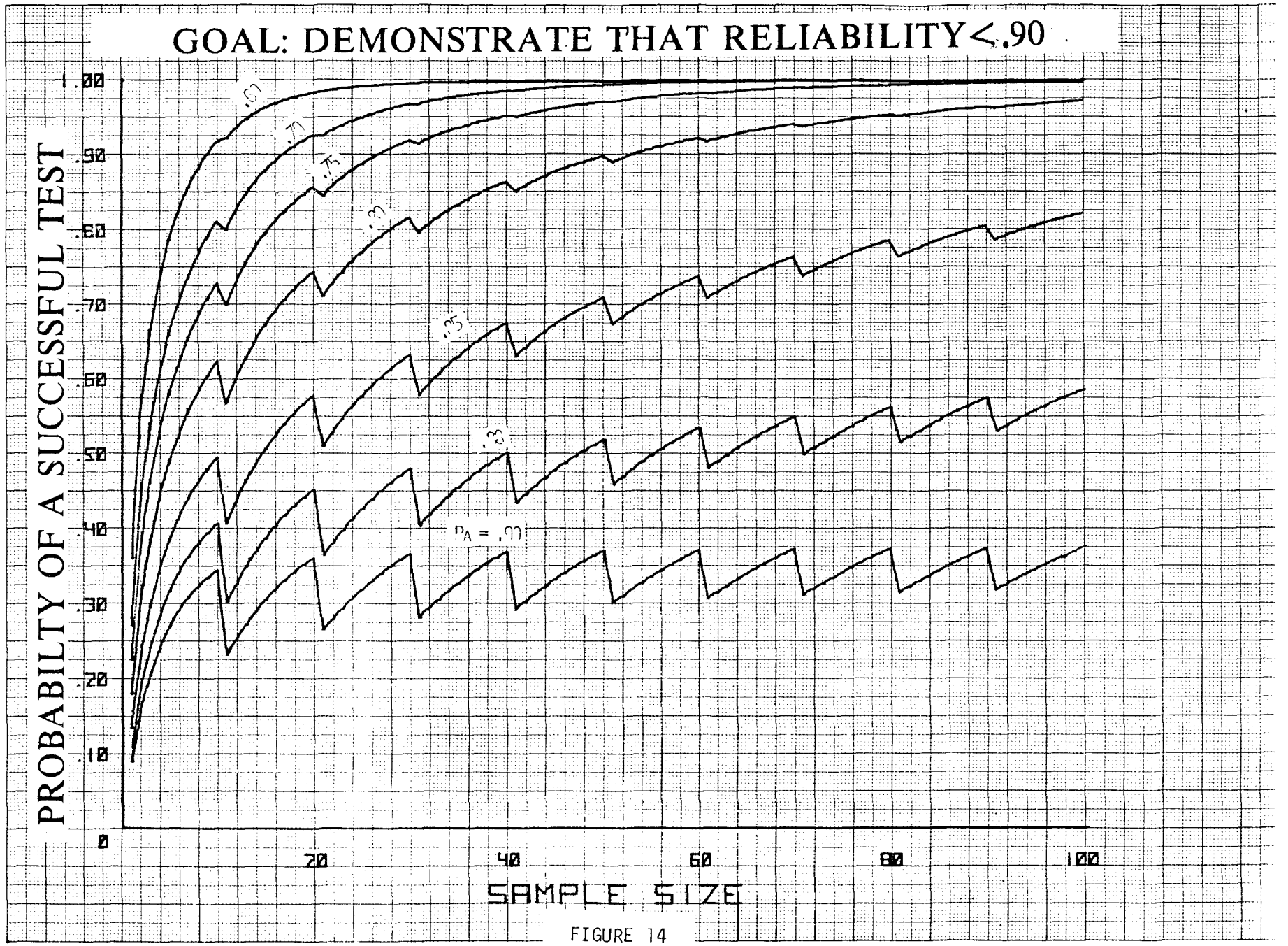

GOAL: DEMONSTRATE THAT RELIABILITY<.90
PROBABILTY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.95
.90
.85
.80
.75
.70
P_A = .90
20
40
60
80
100
SAMPLE SIZE

FIGURE 14

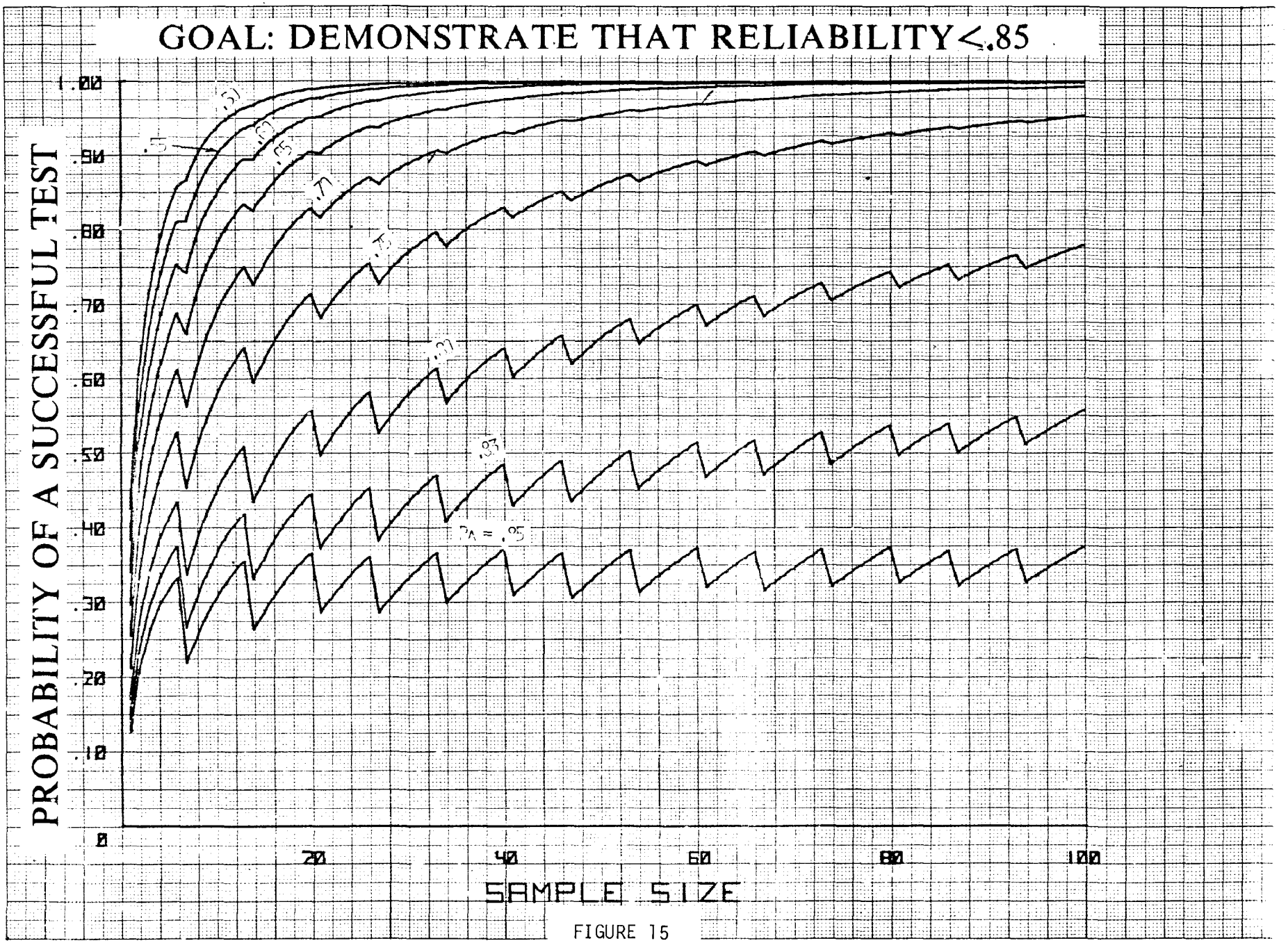

GOAL: DEMONSTRATE THAT RELIABILITY<.85
PROBABILITY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
20
40
60
80
100
SAMPLE SIZE
.97
.95
.93
.9
.85
.8
.75
.7
.65
.6
.55
.5
.45
.4
.35
.3
.25
.2
.15
.1
.05

FIGURE 15

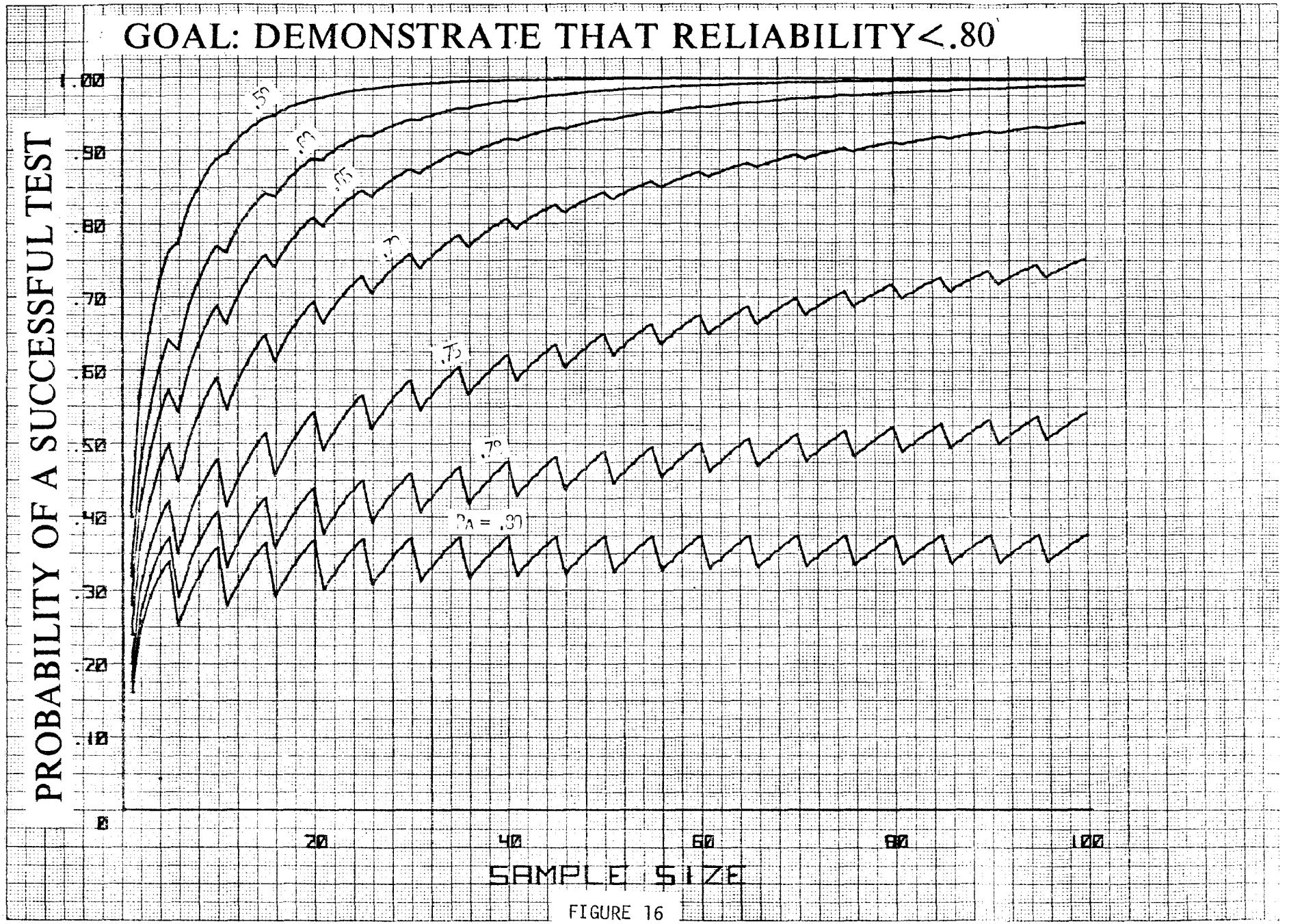

GOAL: DEMONSTRATE THAT RELIABILITY<.80
PROBABILITY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.55
.60
.65
.70
.75
.80
p_A = .30
20
40
60
80
100
SAMPLE SIZE

FIGURE 16

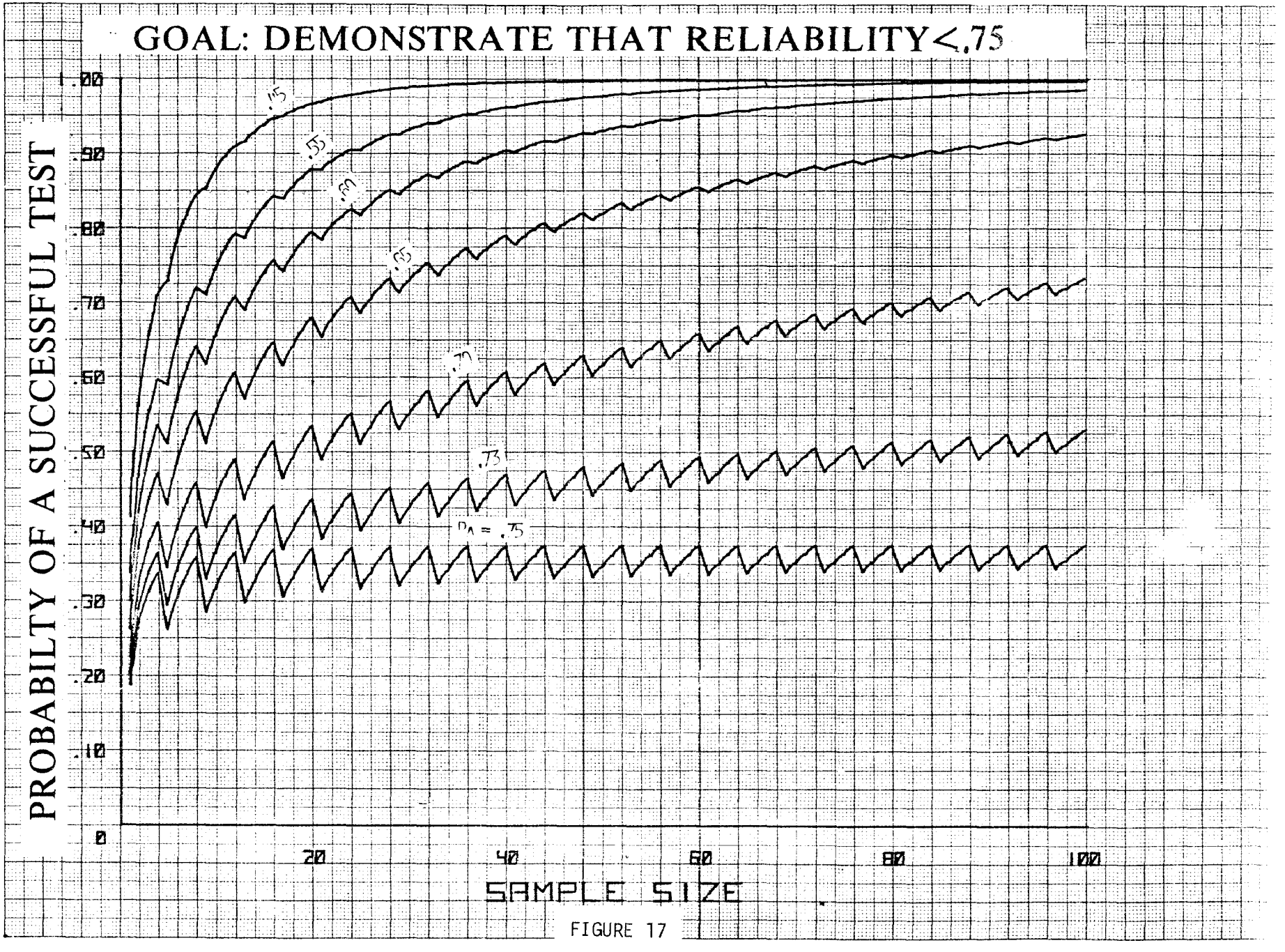

GOAL: DEMONSTRATE THAT RELIABILITY<.75
PROBABILTY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.95
.55
.50
.65
.70
.73
p_A = .75
20
40
60
80
100
SAMPLE SIZE

FIGURE 17

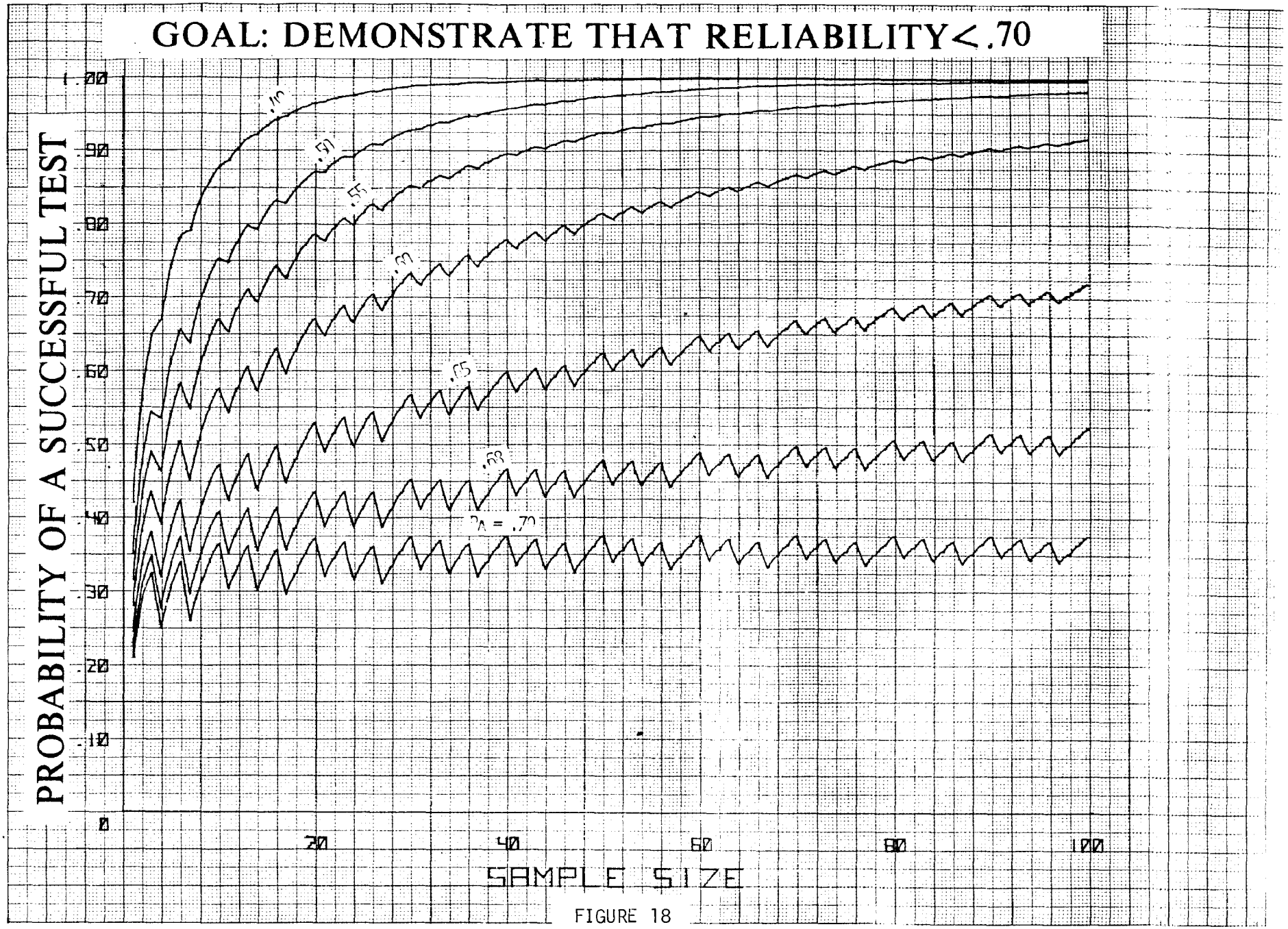

GOAL: DEMONSTRATE THAT RELIABILITY < .70
PROBABILITY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.50
.55
.60
.65
.68
.70
20
40
60
80
100
SAMPLE SIZE

FIGURE 18

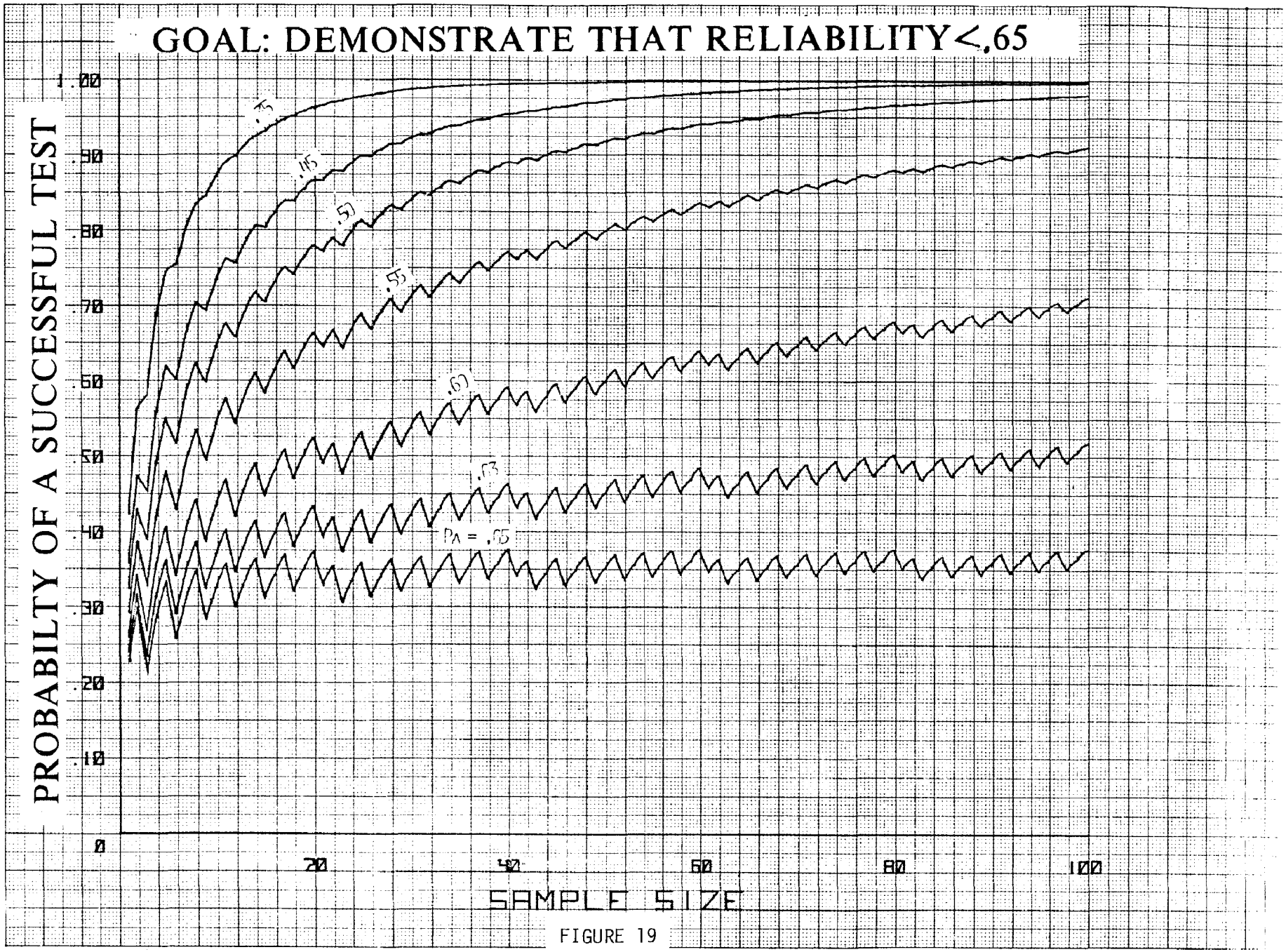

GOAL: DEMONSTRATE THAT RELIABILITY <.65
PROBABILTY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.75
.65
.57
.55
.63
.53
PA = .65
20
40
60
80
100
SAMPLE SIZE

FIGURE 19

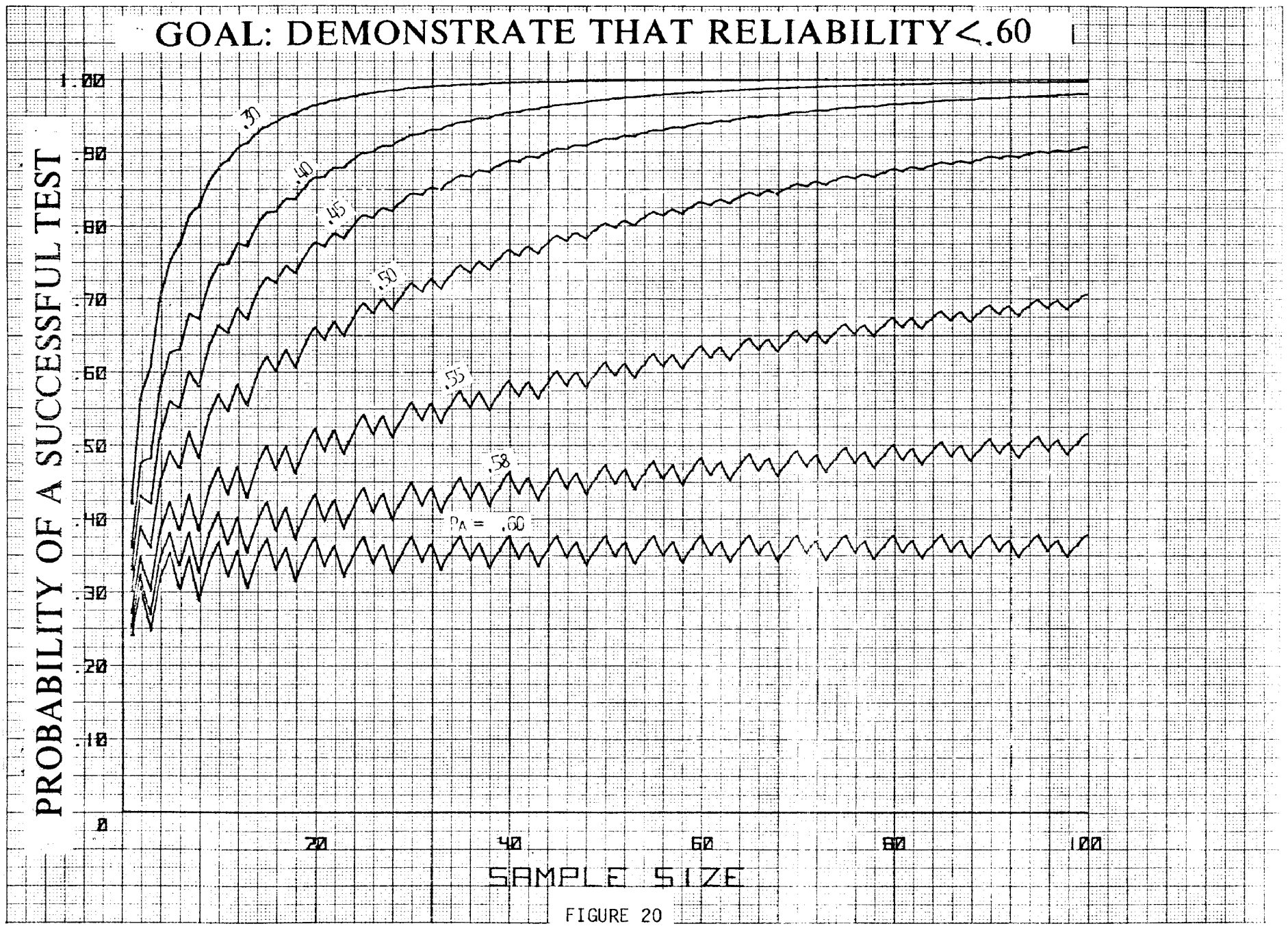

GOAL: DEMONSTRATE THAT RELIABILITY<.60
PROBABILITY OF A SUCCESSFUL TEST
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
.30
.40
.45
.50
.55
.58
P_A = .60
20
40
60
80
100
SAMPLE SIZE

FIGURE 20

DOVAP BEST ESTIMATE OF TRAJECTORY

Robert H. Turner and William S. Agee

US Army White Sands Missile Range, NM 88002

ABSTRACT. The DOVAP Best Estimate of Trajectory (BET) is a new development in trajectory data reduction at WSMR. The DOVAP BET provides the capability to produce high quality trajectory solutions one to three days after a flight test. It is batch processing type of trajectory solution - it estimates all trajectory points simultaneously. This type of solution inherently results in a very large but sparse set of equations. In the DOVAP case this set of equations is block tridiagonal and is solved by an LU matrix decomposition. The solution for the trajectory from the LU decomposition must be iterated since the original DOVAP equations are nonlinear. Numerical procedures, program structure, and data editing procedures used in the DOVAP BET program are described.

1. INTRODUCTION. The DOVAP Best Estimate of Trajectory (BET) program provides the ability to produce high quality trajectory solutions one to three days after a mission. More specifically, a DOVAP BET can be provided in 2-3 work days with normal availability of digitized DOVAP tapes. Upon special request and arrangements by the Range User, a DOVAP BET can be provided in one work day. Real-time digitizing and central recording of DOVAP data would make 1-2 day DOVAP trajectory solutions a standard item.

The DOVAP BET is a batch processor - it estimates the entire trajectory at once. The editing of observations, which takes care of both wild data and bias, is entirely automatic. The DOVAP batch processor,since it solves a set of nonlinear equations, requires a starting guess for the trajectory. The program will accept trajectory guesses from several sources. These guesses may often be several miles in error without affecting the convergence of the batch processor. The guessed trajectory must be reasonably smooth, i.e., it must be free from spikes. Sources of guessed trajectories presently used in the DOVAP BET program are radar, a nominal trajectory or flight path, or a sequence of constant points. Since the input to the DOVAP BET program is a merged data tape, the most common source of a guess trajectory is radar which is inputted from the same merged tape. The use of radar guesses does not slow the production of a DOVAP BET if the central record facility for radar is used. The radar data need not be calibrated for use as a trajectory guess. The DOVAP BET program automatically despikes the radar observations before use in constructing a trajectory guess.

The DOVAP measurements are refraction corrected by the method described in Appendix A. Since the DOVAP tracking ranges are relatively short and the DOVAP BET program automatically rejects DOVAP observations when the tracking elevation is below a preset minimum, an approximation to flat Earth ray tracing equations has been found to provide a sufficiently

accurate elevation refraction correction for use in the method of Appendix A.

2. DOVAP MEASUREMENT MODEL. The DOVAP measurement system is a two way CW doppler system which measures a loop range change or equivalently an average loop range rate of a target between consecutive sampling times t_i and t_{i+1}. The normal sampling interval is 50 msec. A DOVAP instrumentation path consists of a transmitter station and usually ten or more receiver stations. In addition the target being tracked carries a frequency doubling transponder. The DOVAP receiver station receives both the reference frequency from the transmitter and the frequency doubled signal from the target transponder. A frequency comparison of the two signals and the addition of a $5KH_z$ bias signal is performed at the receiver. The resulting signal is converted to digital form by counting the doppler cycles over a 50 msec time interval. The resulting biased doppler measurement is

$$M(t_{i+1}) = 14.72194\ s(t_i,t_{i+1}) + 50278 + N_{i+1} \tag{1}$$

where $s(t_i,t_{i+1})$ is the refraction distorted change in loop range over the 50 msec interval. The loop range is defined as the sum of the range from transmitter to target and the range from target to receiver. N_{i+1} is a measurement error which is assumed to be random with zero mean. Thus, except for refraction, for which a correction is applied, DOVAP measurements are considered to be unbiased. Subtracting the bias 50278 and dividing by 14.72194 yields the modified DOVAP measurement $M'(t_{i+1})$

$$M'(t_{i+1}) = s(t_i,t_{i+1}) + N_{i+1} \tag{2}$$

The correction to this measurement for refraction is discussed in Appendix A. After correcting for refraction we have the measurement model for the α^{th} receiver.

$$z_\alpha(t_i,t_{i+1}) = g_\alpha(x_i,x_{i+1}) + v_\alpha(i+1) \tag{3}$$

where $g_\alpha(x_i,x_{i+1})$ is the loop range change from t_i to t_{i+1} and the arguments x_i and x_{i+1} indicate the dependence on the position coordinates at these times. x_i is the position vector of the target in the launcher coordinate system.

The loop range change is modelled as follows. Let (x_T,y_T,z_T) be the coordinates of the electrical center of the ground transmitter in the launcher coordinate system. Also, let (x_R,y_R,z_R) be the coordinates at the electrical center of the ground receiver antenna. The z coordinates of the surveyed transmitter and receiver sites are modified to obtain z_T and z_R. The amount of modification depends on the type of antenna used. Let (x,y,z) be the launcher coordinates of the target transmitting and receiving antennas. Let R_T be the range from ground transmitter antenna to target receiving antenna and R_R be the range from target transmitting antenna to ground receiver antenna. R_T and R_R are given by

(4) $$R_T(t_i) = [(x(t_i)-x_T)^2+(y(t_i)-y_T)^2+(z(t_i)-z_T)^2]^{\frac{1}{2}}$$

(5) $$R_R(t_i) = [(x(t_i)-x_R)^2+(y(t_i)-y_R)^2+(z(t_i)-z_R)^2]^{\frac{1}{2}}$$

Then the loop range change $g(x_i, x_{i+1})$ is

(6) $$g(x_i, x_{i+1}) = R_T(t_{i+1}) + R_R(t_{i+1}) - R_T(t_i) - R_R(t_i)$$

The partial derivatives required for processing DOVAP observations using the above measurement model are

(7) $$\begin{bmatrix} \frac{\partial g}{\partial x(t_i)} \\ \frac{\partial g}{\partial y(t_i)} \\ \frac{\partial g}{\partial z(t_i)} \end{bmatrix} = \begin{bmatrix} \frac{-(x(t_i)-x_T)}{R_T(t_i)} - \frac{(x(t_i)-x_R)}{R_R(t_i)} \\ \frac{-(y(t_i)-y_T)}{R_T(t_i)} - \frac{(y(t_i)-y_R)}{R_R(t_i)} \\ \frac{-(z(t_i)-z_T)}{R_T(t_i)} - \frac{(z(t_i)-z_R)}{R_R(t_i)} \end{bmatrix}$$

(8) $$\begin{bmatrix} \frac{\partial g}{\partial x(t_{i+1})} \\ \frac{\partial g}{\partial y(t_{i+1})} \\ \frac{\partial g}{\partial z(t_{i+1})} \end{bmatrix} = \begin{bmatrix} \frac{(x(t_{i+1})-x_T)}{R_T(t_{i+1})} + \frac{(x(t_{i+1})-x_R)}{R_R(t_{i+1})} \\ \frac{(y(t_{i+1})-y_T)}{R_T(t_{i+1})} + \frac{(y(t_{i+1})-y_R)}{R_R(t_{i+1})} \\ \frac{(z(t_{i+1})-z_T)}{R_T(t_{i+1})} + \frac{(z(t_{i+1})-z_R)}{R_R(t_{i+1})} \end{bmatrix}$$

3. THE DOVAP BATCH PROCESSOR. Given a set of N time points $t_i, i=1,N$ along a trajectory and a set of N-1 loop range change observations $Z_\alpha(i)$, $i=2,N$ for M DOVAP receivers $\alpha=1,M$, we want to estimate the position vector x_i for each of the trajectory time points, t_i. Recall the DOVAP observation model given by (2-3)

$$Z_\alpha(t_{i-1},t_i) = g_\alpha(x_{i-1},x_i) + v_\alpha(i) \tag{9}$$

The measurement error $v_\alpha(i)$ is assumed to have zero mean and variance $R_\alpha(i)$.

The DOVAP batch processor minimizes the weighted sum of squares

$$Q(x_1,x_2,\ \ldots,x_N) = \sum_{i=2}^{N} \sum_{\alpha=1}^{M} \frac{1}{R_\alpha(i)} \left(Z_\alpha(t_{i-1},t_i)-g_\alpha(x_{i-1},x_i)\right)^2 \tag{10}$$

Taking derivatives of Q with respect to $x_1,x_2,\ \ldots,x_N$ and equating to zero results in the equations

$$\sum_{\alpha=1}^{M} \frac{1}{R_\alpha(2)}\ G_{\alpha 1}^T(x_1)\left(Z_\alpha(t_1,t_2)-g_\alpha(x_1,x_2)\right)=0 \tag{11}$$

$$\sum_{\alpha=1}^{M} \frac{1}{R_\alpha(j+1)}\ G_{\alpha 1}^T(x_j)\left(Z_\alpha(t_j,t_{j+1})-g_\alpha(x_j,x_{j+1})\right)+ \frac{1}{R_\alpha(j)}\ G_{\alpha 2}^T(x_j) \tag{12}$$

$$\left(Z_\alpha(t_{j-1},t_j)-g_\alpha(x_{j-1},x_j)\right) = 0 \qquad j=2,N-1$$

$$\sum_{\alpha=1}^{M} \frac{1}{R_\alpha(N)}\ G_{\alpha 2}^T(x_N)\left(Z_\alpha(t_{N-1},t_N)-g_\alpha(x_{N-1},x_N)\right)=0 \tag{13}$$

where

$$G_{\alpha 1}^T(x_K) = \frac{\partial g_\alpha(x_K,x_{K+1})}{\partial x_K} \tag{14}$$

$$G_{\alpha 2}^T(x_{K+1}) = \frac{\partial g_\alpha(x_k,x_{k+1})}{\partial x_{k+1}} \tag{15}$$

In order to solve the above system of 3N nonlinear equations, it is necessary to linearize these equations about a set of points $x_1^o, x_2^o, \ldots, x_N^o$. The set of points may be obtained from a number of sources and need not be very close to the solution. The linearization results in a set of linear equations to be solved for $\delta x_i = x_i - x_i^o$. The nonlinear equations are then relinearized about the new trial solution $x_i^1 = x_i^o + \delta x_i$. This process is repeated until the solution converges, i.e., $|\delta x_i| \to 0$. The following development of the solution of the linearized equations indicates only the first iteration in the process of solving the set of nonlinear equations. Each succeeding iteration in the process is identical to the first one with x_i^o replaced by $x_i^o + \delta x_i$. Linearization of (11-13) results in

$$A_1 \delta x_1 + A_{12} \delta x_2 = y(1) \tag{16}$$

$$A_{j-1,j}^T \delta x_{j-1} + A_j \delta x_j + A_{j,j+1} \delta x_{j+1} = y(j), \quad j=2, N-1 \tag{17}$$

$$A_{N-1,N}^T \delta x_{N-1} + A_N \delta x_N = y(N) \tag{18}$$

where

$$A_1 = \sum_{\alpha=1}^{M} \frac{1}{R_\alpha(2)} G_{\alpha 1}^T (x_1^o)\, G_{\alpha 1}(x_1^o) \tag{19}$$

$$A_j = \sum_{\alpha=1}^{M} \frac{1}{R_{\alpha(j+1)}} G_{\alpha 1}^T (x_j^o) G_{\alpha 1}(x_j^o) + \frac{1}{R_{\alpha(j)}} G_{\alpha 2}^T (x_j^o)\, G_{\alpha 2}(x_j^o) \tag{20}$$

$$A_N = \sum_{\alpha=1}^{M} \frac{1}{R_\alpha(N)} G_{\alpha 2}^T (x_N^o)\, G_{\alpha 2}(x_N^o) \tag{21}$$

$$A_{K,K+1} = \sum_{\alpha=1}^{M} \frac{1}{R_\alpha(K+1)} G_{\alpha 1}^T (x_K^o)\, G_{\alpha 2}\, (x_{K+1}^o) \tag{22}$$

$$y(1) = \sum_{\alpha=1}^{M} \frac{1}{R_\alpha(2)} G_{\alpha 1}^T (x_1^o)\, r_\alpha(2) \tag{23}$$

$$y(j) = \sum_{\alpha=1}^{M} \frac{1}{R_\alpha(j+1)} G_{\alpha 1}^T (x_j^o)\, r_\alpha(j+1) + \frac{1}{R_\alpha(j)} G_{\alpha 2}^T (x_j^o) r_\alpha(j) \tag{24}$$

$$y(N) = \sum_{\alpha=1}^{M} \frac{1}{R_\alpha(N)} G_{\alpha 2}^T (x_N^o)\, r_\alpha(N) \tag{25}$$

r (j) is the residual

$$r(j) = Z(t_{j-1}, t_j) - g(x_{j-1}^{o}, x_j^{o}) \tag{26}$$

The system of linear equations (16-18) can be written in the block tridiagonal matrix form

$$\begin{bmatrix} A_1 & A_{12} & 0 & 0 & \cdots & 0 \\ A_{12}^T & A_2 & A_{23} & 0 & \cdots & 0 \\ 0 & A_{23}^T & A_3 & A_{34} & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & \cdots & A_{N-2,N-1}^T & A_{N-1} & A_{N-1,N} \\ 0 & 0 & \cdots & 0 & A_{N-1,N}^T & A_N \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \\ \delta x_3 \\ \vdots \\ \delta x_{N-1} \\ \delta x_N \end{bmatrix} = \begin{bmatrix} y(1) \\ y(2) \\ y(3) \\ \vdots \\ y(N-1) \\ y(N) \end{bmatrix} \tag{27}$$

This set of equations to be solved for the δx's is of very large dimension since N, the number of trajectory points to be estimated, is typically in the range 50-300. In addition these equations must be solved many times in the iteration process to solve the nonlinear set of equations in (11-13). Fortunately, there is a very simple form solving the block tridiagonal system equations of (27).

Let A be the above block tridiagonal matrix. Since A is a least squares matrix, it is symmetric and positive definite. A standard theorem in matrix theory is that for any symmetric positive definite A there exists a non-singular, lower triangular matrix L such that

$$A = LL^T \tag{28}$$

In the case of our block tridiagonal matrix A the matrix L has an especially simple form. In this case we can write the decomposition (28) as

(29)
$$\begin{bmatrix} L_1 & & & \\ L_{21} & L_2 & & \\ 0 & L_{32} & L_3 & \\ & 0 & & \\ \vdots & \vdots & & \\ 0 & 0 & L_{N,N-1} & L_N \end{bmatrix} \begin{bmatrix} L_1^T & L_{21}^T & 0 & \cdots & 0 \\ & L_2^T & L_{32}^T & \cdots & 0 \\ & & L_3^T & & \vdots \\ & & & & L_{N,N-1}^T \\ & & & & L_N^T \end{bmatrix}$$

All block elements of the above block triangular matrices are 3x3. Equating block elements of A with the corresponding block elements of LL^T, we find the following set of equations.

(30) $$A_1 = L_1 L_1^T$$

(31) $$A_{12} = L_1 L_{21}^T$$

(32) $$A_i = L_{i,i-1} L_{i,i-1}^T + L_i L_i^T \qquad i=2,N$$

(33) $$A_{i,i+1} = L_i L_{i+1,i}^T \qquad i=2,N$$

Since the matrices L_i are lower triangular a Cholesky decomposition algorithm is used to solve (30) and (32) for L_i. Also, since the L_i are lower triangular, (31) and (33) defining the off diagonal blocks $L_{i,i-1}$ are solved as triangular systems of equations for the columns of $L_{i,i-1}$.

The square root decomposition of A allows us to replace the set of equations $A\delta x=y$ given by (27) by two sets of equations

(34) $$L\, y' = y$$

(35) $$L^T \delta x = y'$$

The lower triangular system of equations in (34) may be written in terms of the 3x3 set of block elements as

(36) $$L_1 y_1' = y(1)$$

(37) $$L_{i+1} y_{i+1}' = y_{(i+1)} - L_{i+1,i} y_i'$$

This set of 3x3 lower triangular equations is solved sequentially for the vectors y_i'. Similarly, the upper triangular system of equations in (35) can be written in terms of the 3x3 set of block elements as

$$L_N^T \, \delta x_N = y_N' \tag{38}$$

$$L_{i-1}^T \, \delta x_{i-1} = y_{i-1}' - L_{i,i-1}^T \, \delta x_i \tag{39}$$

This set of equations is then solved sequentially for the vectors δx_i.

From the theory of weighted least sequares we know that the covariance of the vector estimate of x obtained on the final iteration is approximately A^{-1} obtained on the final iteration. Due to the method of solution of the equations, A^{-1} is not available. However, it is relatively easy to obtain the covariance of the individual vector estimates x_i from the block elements of the square root matrix L. A^{-1} is given by

$$A^{-1} = L^{-T} L^{-1} \tag{40}$$

It is easily determined that the block elements of L^{-1} are given by

$$L_{ii}^{-1} = L_i^{-1} \qquad i=1,N \tag{41}$$

$$L_{i,j-1}^{-1} = -L_{ij}^{-1} \, L_{j,j-1} \, L_{j-1}^{-1} \qquad j=2,i \tag{42}$$

The covariance matrices of the estimates x_i are

$$\text{cov}(x_i) = A_{ii}^{-1} = \sum_{j=i}^{N} L_{ij}^{-T} \, L_{ji}^{-1} \tag{43}$$

A computing formula for these 3x3 covariance matrices is obtained by substituting the recursive relations for L_{ij}^{-1} given by (42) into (43). Thus

$$A_{i-1,i-1}^{-1} = \sum_{j=i-1}^{N} L_{i-1,j}^{-T} \, L_{j,i-1}^{-1} \tag{44}$$

$$A_{i-1,i-1}^{-1} = \sum_{j=i}^{N} L_{i-1,j}^{-T} \, L_{j,i-1}^{-1} + L_{i-1}^{-T} \, L_{i-1}^{-1} \tag{45}$$

Using (42)

$$A_{i-1,i-1}^{-1} = \sum_{j=i}^{N} (L_{i-1}^{-T} L_{i,i-1}^{T}) \, L_{ij}^{-T} \, L_{ji}^{-1} \, (L_{i,i-1} L_{i-1}^{-1}) + L_{i-1}^{-T} \, L_{i-1}^{-1} \tag{46}$$

$$A^{-1}_{i-1,i-1} = L^{-T}_{i-1}\,[L^{T}_{i,i-1}\,A^{-1}_{ii}\,L_{i,i-1} + I]\,L^{-1}_{i-1} \tag{47}$$

Obviously,

$$A^{-1}_{NN} = L^{-T}_{N}\,L^{-1}_{N} \tag{48}$$

4. FRONT END EDIT. Before using the DOVAP measurements in the batch processor, wild observations are detected and deleted and the variances $R_\alpha(j)$, which are used in forming weights for the batch processor, are computed.

For the α^{th} receiver consider the set of observation $\{Z_\alpha(t_j,t_{j+1})\}$ on the interval $[(t_i,t_{i+N_p})]$. If no observations have been drop locked for this receiver, there will be N_p equally spaced observations in this interval. Let N_α be the actual number of observations in the interval. A quadratic curve is fitted to this set of N_α observations. Let $\hat{Z}_\alpha(t_{j-1},t_j)$ denote the value of the curve at t_j and let $\tilde{Z}_\alpha(t_{j-1},t_j) = Z_\alpha(t_{j-1},t_j) - \hat{Z}_\alpha(t_{j-1},t_j)$ be the residual from the curve fit. The set of residuals on the interval $[(t_i,t_{i+Np})]$ are tested for spurious observations using the sample skewness coefficient $\sqrt{b_1}$ and the sample Kurtosis coefficient b_2. These statistics are given by

$$\sqrt{b_1} = \frac{\sqrt{N_a}\ \sum_i (\tilde{Z}_\alpha(t_{i-1},t_i)-\mu_\alpha)^3}{(N_\alpha-1)\ s_\alpha{}^{3}/_{2}} \tag{49}$$

$$b_2 = \frac{N_\alpha\ \sum_i (\tilde{Z}_\alpha(t_{i-1},t_i)-\mu_\alpha)^4}{(N_\alpha-1)\ s_\alpha{}^4} \tag{50}$$

where μ_α and $s_\alpha{}^2$ are the sample mean and variance, respectively. If either $|\sqrt{b_1}|$ or b_2 exceed values for the 5% significance level of these statistics, the observation farthest from the mean is deleted. A curve is fitted to the remaining observations in the interval and the test for spurious observations is again applied. This sequence of fitting and testing is repeated until $|\sqrt{b_1}|$ and b_2 are less than the values corresponding to the 5% significance levels or until less than $.75N_p$ observations remain in the interval. When no more observations are to be deleted, the points which were deleted are replaced by the corresponding values of the final curve fit. In addition all points which are originally missing from the interval, i.e., the drop locked points, are replaced by their corresponding values from the curve fit. If more than $.25N_p$ of the observations in the interval require replacement, all observations in the interval are deleted and not replaced.

The interval (t_i, t_i, N_p) is chosen so that it has an integral number of subintervals each having L_p observations. The observations in each of these subintervals are summed to provide loop range change measurements to the batch processor which are $.05L_p$ sec apart. The variance of these observations, which are used in the batch processor, are computed as $R_\alpha(i)=L_p s_\alpha^2(i)$ where s_α^2 is the sample variance computed from the curve fit on the interval (t_i, t_{i+N_p}). These will be N_p/L_p consecutive variances $R_\alpha(i)$ which will be equal for the α^{th} receiver. The above editing procedure is applied to each receiver and to a set of intervals which cover the entire trajectory.

Replacement of wild observations is not in itself a desirable procedure and is certainly not necessary. However, in the DOVAP batch processor this replacement procedure greatly simplifies the operational logic of the computer program. By being conservative in the replacement of observations and realizing the replacement can be dangerous, no apparent difficulties have been encountered.

5. POINT EDIT AND RECEIVER EDIT. Although the front end edit has deleted most the spurious observations, some may still remain. More importantly, a receivers observations may be biased either over a small portion of the trajectory or over the entire trajectory. The point edit and receiver edit are designed to delete observations in these categories.

Following the front end edit and iteration of the batch processor to convergence, the point edit procedure is applied to each receiver observation used in the batch processor. At each trajectory time, t_i, the normalized residuals

$$r^*_\alpha(i) = \frac{r_\alpha(i)}{\sqrt{R_\alpha(i)}} \tag{51}$$

are examined for each receiver α used in the batch processing. The residual $r_\alpha(i)$ is defined by (26) and is the residual computed in the last iteration of the batch processor. If $|r^*_\alpha(i)| \geq 5$, corresponding observation is deleted permanently from the solution. The observation corresponding to the largest of the $|r^*_\alpha(i)|$ such that $3 \leq |r^*_\alpha(i)| < 5$ is deleted temporarily. After examining residuals for the entire trajectory and deleting the ones dictated by the above tests, the batch processor is reiterated to convergence. Additional observations are then removed according to the above criteria. The iteration and point edit cycle is repeated until $|r^*_\alpha(i)| < 3$ for all remaining residuals.

When the above iteration and point edit cycle is completed, a receiver edit identifies receivers suspected of being biased. The receiver bias edit computes the statistic

$$m_\alpha = \frac{1}{\sqrt{M_a}} \sum r^*_\alpha(i) \tag{52}$$

for each receiver. The sum in the above equation is over all residuals for the α^{th} receiver which were not deleted, either temporarily or permanently by the point edit. M_α is the number of terms in this sum. If $m = \max_\alpha |m_\alpha| \geq 3$, the receiver corresponding to the value m is considered to be biased and is deleted from the batch processing solution. The iteration and point edit cycle is then repeated without this receiver. All residuals which were temporarily deleted on the previous point edit cycle are restored to the solution. The receiver edit is repeated following an iteration and point edit cycle until no additional receivers are suspected of being biased.

6. TRAJECTORY FILLING. The DOVAP batch processing solution produces a trajectory position solution at time points spaced $.05L_p$ sec apart. The number L_p is usually chosen so that these time points are 1-2 sec apart. It is obvious that a 1-2 sec interval between trajectory position solutions is unacceptable for most purposes. This is particularly true if velocity and acceleration states are to be derived from the position solution. Although we could set $L_p=1$ so that a batch processor solution is computed at every DOVAP sample time, this is undesirable since the CPU time and mass storage used by the DOVAP BET program is greatly increased. A procedure described below, which we call trajectory filling, is used to add position solutions between the batch processor position solutions. Experimental results show that there is very little difference between the position solutions obtained by trajectory filling and those obtained by setting $L_p=1$.

Let the DOVAP sample times between batch processor times t_i and t_{i+1} be denoted by $t_{i,1},t_{i,2},\ldots,t_{i,Lp-1}$ and the corresponding trajectory position vectors by $x_{i,1},x_{i,2},\ldots,x_{i,Lp-1}$. The least squares position solution for computing $x_{i,j+1}$, given $x_{i,j}$ is

$$x_{i,j+1}=x_{i,j}+M^{-1}\sum_{\alpha}\frac{1}{s_\alpha^2(i)}\,G_{\alpha 2}^T(x_{i,j})\,r_\alpha(t_{i,j},t_{i,j+1}) \tag{53}$$

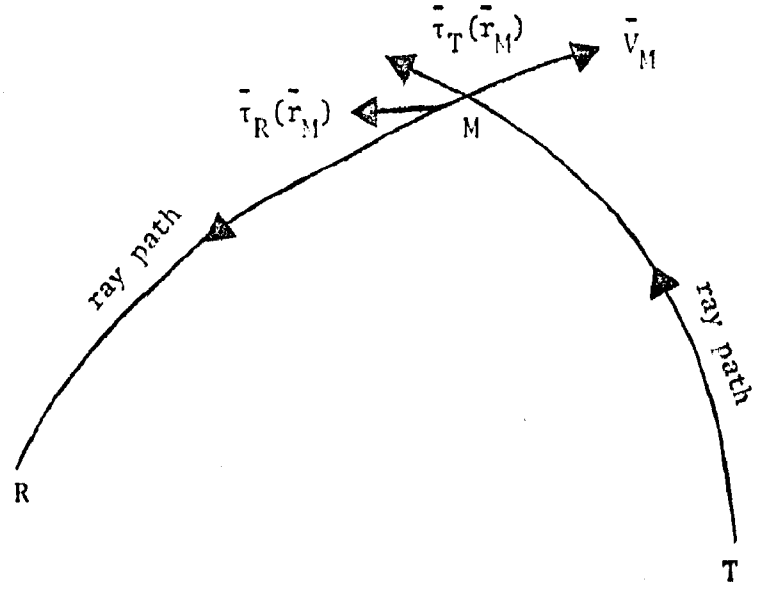
where

$$M=\sum_{\alpha}\frac{1}{s_\alpha^2(i)}\,G_{\alpha 2}^T(x_{i,j})\,G_{\alpha 2}(x_{i,j}) \tag{54}$$

The sums in the above two equations are over all receivers whose observations were in the final batch processor iteration for the time interval t_i to t_{i+1}. The quantity $s_\alpha^2(i)$ is the sample variance for the interval t_i to t_{i+1} which was computed in the front end edit.

DOVAP REFRACTION CORRECTION. The refraction correction applied to DOVAP is considerably more complex than that normally applied. Since the ranges involved in DOVAP tracking are relatively short, a flat Earth ray tracing procedure can be used. Also, since DOVAP data is normally used only when the elevation angles are sufficiently large, say greater that 10°; an approximation to the ray trace procedure is sufficiently accurate.

In the figure below the DOVAP transmitter is located at T and the receiver at R. The target is at M travelling with velocity $\bar{V}_M$.



$\bar{\tau}_T(\bar{r}_M)$ and $\bar{\tau}_R(\bar{r}_M)$ are unit vectors tangent at the target to the ray paths from transmitter to target and target to receiver, respectively.

For an isotropic medium and for targets with speed $|\bar{V}_M| << c$, the doppler frequency, f_D, at the receiver can be expressed in terms of the phase path length L.

$$f_D = \frac{-2fo}{c} \frac{dL}{dt} \tag{55}$$

where fo is the transmitted frequency. The phase path length for the above figure is

$$L = \int_T^M n(\bar{r})ds + \int_M^R n(\bar{r})ds \tag{56}$$

where $n(\bar{r})$ is the index of refraction at a point on the ray path. The first term in (56) is the phase path length from transmitter to target and the second term is the phase path length from target to receiver. The integration is to be carried out along the ray path.

The element of path length ds may be written as

$$ds = \bar{\tau}(\bar{r}) \cdot d\bar{r} \tag{57}$$

where $\bar{\tau}(\bar{r})$ is a unit tangent vector to the ray path at point $\bar{r}$.

The path length can then be written as

$$L = \int_{\bar{r}_T}^{\bar{r}_M(t)} n(\bar{r})\bar{\tau}(\bar{r}) \cdot d\bar{r} + \int_{\bar{r}_M(t)}^{\bar{r}_R} n(\bar{r})\bar{\tau}(\bar{r}) \cdot d\bar{r} \tag{58}$$

Differentiating (58) with respect to t and ignoring $\frac{\partial n}{\partial t}$

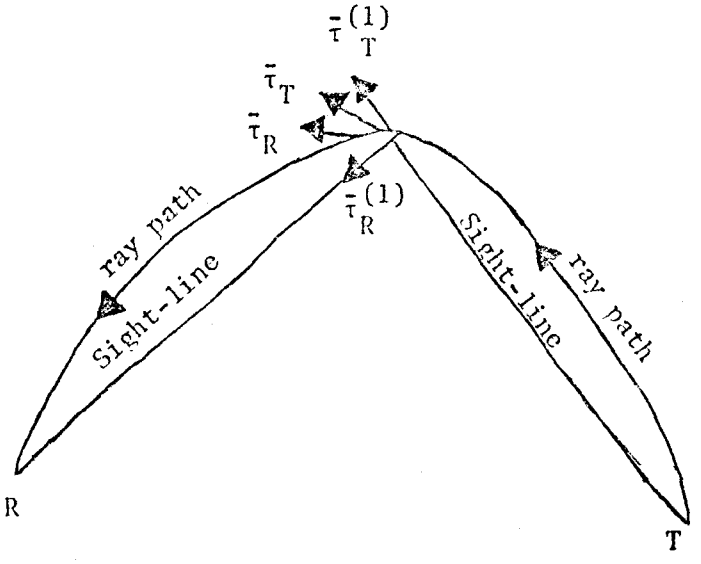
$$\frac{dL}{dt} = n(\bar{r}_{M-})\bar{\tau}(\bar{r}_{M-}) \cdot \frac{d\bar{r}_{M-}}{dt} - n(\bar{r}_{M+})\bar{\tau}(\bar{r}_{M+}) \cdot \frac{d\bar{r}_{M+}}{dt} \tag{59}$$

where the - and + refer to the transmitter and receiver sides of the target, respectively. (59) can be rewritten as

$$\frac{dL}{dt} = n(\bar{r}_M)(\bar{\tau}_T(\bar{r}_M) - \bar{\tau}_R(\bar{r}_M)) \cdot \bar{V}_M \tag{60}$$

(60) shows that the received doppler signal is dependent on the refraction effects of the atmosphere in two ways. First, note that the received doppler is proportional to $n(\bar{r}_M)$, the coefficient of refraction at the target. Second, the received doppler is not quite proportional to the radial velocity as is assumed since the unit vectors $\bar{\tau}_R$ and $\bar{\tau}_T$ are not along the sight line vectors.

Let $\bar{\tau}_T^{(1)}(\bar{r}_M)$ and $\bar{\tau}_R^{(1)}(\bar{r}_M)$ be unit vectors along the sight lines from transmitter to target and target to receiver, respectively. The geometric situation is shown below.



The unit sight-line vectors can be written as

$$\bar{\tau}_T^{(1)} = \frac{(\bar{r}_M - \bar{r}_T)}{|\bar{r}_M - \bar{r}_T|} \tag{61}$$

$$\bar{\tau}_R^{(1)} = -\frac{(\bar{r}_M - \bar{r}_R)}{|\bar{r}_M - \bar{r}_R|} \tag{62}$$

where $\bar{r}_T$ and $\bar{r}_R$ are position vectors of the transmitter and receiver, respectively. Let v_{doBs} be the observed doppler velocity given by (60) and let v_d be the doppler velocity if there were no refraction effects. v_d is given by

$$v_d = (\bar{\tau}_T^{(1)} - \bar{\tau}_R^{(1)}) \cdot \bar{V}_M \tag{63}$$

We can write v_d in terms of v_{doBs} as

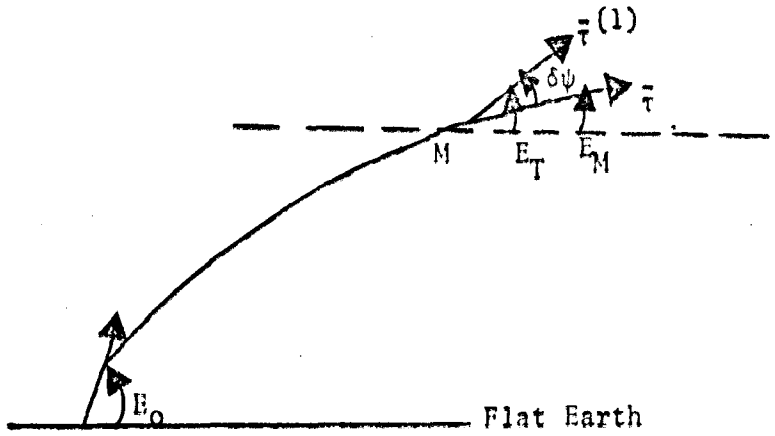
(64) $$v_d = \frac{v_{doBs}}{n(\bar{r}_M)} - \Delta v_d$$

where

(65) $$\Delta v_d = (\bar{\tau}_T - \bar{\tau}_T^{(1)}) \cdot \bar{V}_M - (\bar{\tau}_R - \bar{\tau}_R^{(1)}) \cdot \bar{V}_M$$

In order to correct for refraction in DOVAP or equivalently to correct v_{doBs} to v_d (64) shows that we first divide v_{doBs} by the coefficient of refraction at the target and then substract Δv_d.

In order to compute $(\bar{\tau} - \bar{\tau}^{(1)}) \cdot \bar{V}_M$ needed in the computation of Δv_d, consider one leg of the DOVAP propagation path shown below



E_o is the observed elevation angle and E_T is the true elevation. The elevation refraction correction is $\Delta E = E_T - E_o$. The angle $\delta\psi$ is given by

(66) $$\delta\psi = E_T - E_M$$

(67) $$\delta\psi = (E_T - E_o) - (E_M - E_o)$$

(68) $$\delta\psi = \Delta E - (E_M - E_o)$$

denoting $(E_M - E_o)$ by ΔE_M

$$\delta\psi = \Delta E - \Delta E_M \tag{69}$$

Both ΔE and ΔE_M are available from a ray trace.

Let $\bar{s}_T>$ be an orthonormal set of base vectors with $\bar{s}_{T1}$ along the sight line vector from transmitter to target and $\bar{s}_T$ in the vertical plane containing $\bar{\tau}_T^{(1)}$. In terms of these base vectors $\bar{\tau}_T^{(1)}$ and $\bar{\tau}_T$ can be written as

$$\tau_T^{(1)} = [1 \quad 0 \quad 0] \begin{bmatrix} \bar{s}_{T1} \\ \bar{s}_{T2} \\ \bar{s}_{T3} \end{bmatrix} \tag{70}$$

$$\bar{\tau}_T = [\cos\delta\psi_T \quad 0 \quad -\sin\delta\psi_T] \begin{bmatrix} \bar{s}_{T1} \\ \bar{s}_{T2} \\ \bar{s}_{T3} \end{bmatrix} \tag{71}$$

Since $\delta\psi_T$ is a small angle (71) can be approximated as

$$\bar{\tau}_T \approx [1 \quad 0 \quad -\delta\psi_T] \begin{bmatrix} \bar{s}_{T1} \\ \bar{s}_{T2} \\ \bar{s}_{T3} \end{bmatrix} \tag{72}$$

The velocity vector $\bar{V}_M$ can also be represented in the $\bar{s}_T>$ basis as

$$\bar{V}_M = <\bar{s}_T \; V_M^S> \tag{73}$$

where

$$V_M^S> = M_{s_T L} \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} \tag{74}$$

$M_{s_T L}$ is the rotation matrix from the launcher coordinate system to the sight line coordinate system defined by $\bar{s}_T >$. $\dot{x},\dot{y}$, and $\dot{z}$ are the coordinates of the velocity vector in the launcher coordinate system. The matrix $M_{s_T L}$ may be computed by rotating thru an azimuth angle

$$\alpha_T = \tan^{-1} \frac{x-x_T}{y-y_T} \tag{75}$$

and then thru an elevation angle

$$\varepsilon_T = \tan^{-1} \frac{z-z_T}{[(x-x_T)^2+(y-y_T)^2]^{1/2}} \tag{76}$$

In terms of these angles $M_{s_T L}$ is

$$\begin{bmatrix} \cos\alpha_T & -\sin\alpha_T & 0 \\ \cos\varepsilon_T\sin\alpha_T & \cos\varepsilon_T\cos\alpha_T & \sin\varepsilon_T \\ -\sin\varepsilon_T\sin\alpha_T & -\sin\varepsilon_T\cos\alpha_T & \cos\varepsilon_T \end{bmatrix} \tag{77}$$

Using (70),(72),(74), and (77) the contribution of the transmitter leg of the propagation path to the refraction correction Δv_d is

$$(\bar{\tau}_T-\bar{\tau}_T^{(1)}) \cdot \bar{V}_M \approx \delta\psi_T(-\dot{x}\sin\varepsilon_T\sin\alpha_T-\dot{y}\sin\varepsilon_T\cos\alpha_T+\dot{z}\cos\varepsilon_T) \tag{78}$$

A similar equation holds for the receiver leg of the ray path

$$-(\bar{\tau}_R-\bar{\tau}_R^{(1)}) \cdot \bar{V}_M \approx -\delta\psi_R(-\dot{x}\sin\varepsilon_R\sin\alpha_R-\dot{y}\sin\varepsilon_R\cos\alpha_R+\dot{z}\cos\varepsilon_R) \tag{79}$$

The correction Δv_d given by (65) can now be computed.

OPTIMAL DOVAP INSTRUMENTATION PLANNING

William S. Agee and Jerry L. Meyer

US Army White Sands Missile Range, NM 88002

ABSTRACT. A DOVAP instrumentation planning procedure has been developed which selects a nearly geometrically optimal set of M receiver sites from over 600 sites presently available at WSMR. The criterion used for selection of the sites is the minimize

$$\sum_{i=1}^{N} w_i \, \mathrm{tr} V_i$$

where Vi are the 3×3 diagonal blocks of the 3N×3N error covariance matrix V which would result if the data from the receivers were batch processed to obtain vector position estimates x_i, i=1,N for N points entirely covering a nominal trajectory. The $w_i>0$ are weights used to denote the relative importance of precise estimates at each of the N trajectory points. The final output of the selection procedure is a list of the M receiver sites, the trajectory segments in which each of these receivers should produce usable doppler data, the geometrical covariance for each trajectory point, and a computer plot of the receiver sites along with the ground track of the nominal flight path.

1. INTRODUCTION. The DOVAP instrumentation planning procedure described in this report is designed to assist in the development of a DOVAP Instrumentation Plan (IP) for a given nominal trajectory or flight path. The technique produces a locally optimal selection of M existing receiver sites based on geometrical considerations. Since other significant factors such as multipath, power requirements, site accessability, etc. are also important in the preparation of a DOVAP instrumentation plan, consideration of the factors may require alteration of the geometrically optimal instrumentation plan. For this reason the computer program also has a compare mode to determine the geometrical degradation between the optimal plan and the altered plan.

Given a nominal flight path specified by the position vectors x_i, i=1,N along the path, the receiver selection process minimizes

$$\sum_{i=1}^{N} w_i \, \mathrm{tr} \, \mathrm{cov}(x_i)$$

where $\mathrm{cov}(x_i)$ is the 3×3 covariance for an estimate of x_i, which would result from estimating the trajectory with the DOVAP BET program [1]. The w_i are positive weights used to emphasize some trajectory points

more than others. The most common choice for w_i is $w_i=1$ for all i, i.e. each trajectory point has equal weight in the receiver selection process. $cov(x_i)$ used in the criterion function is a purely geometric covariance, i.e., all receiver noise covariances are unity.

Selection of the best possible set of M DOVAP receivers from the 650 receiver sites presently available at WSMR is a very large combinatorial programming problem. Enumeration and examination of all possible sets of M receivers for a given flight path is obviously an impossible computing problem. Several methods for the solution of integer and combinatorial programming problems are available in the literature on Operations Research. However, the author has not been able to formulate the present problem in a way which would be amenable to solution by these methods. Rather than pursuing the globally optimal set of M DOVAP receivers, we will find a locally optimal set of receivers based on a plan improvement algorithm. Starting with some arbitrary initial IP having M receivers, we make successive interchanges between receivers presently in the IP and those outside the plan until no further improvements are possible. At each stage of the improvement algorithm the interchange is made which results in the greatest improvement to the IP according to the minimization criterion. The resulting plan is only locally optimal in the sense that it is dependent on the initial plan with which the algorithm started. Starting with a different initial IP we might achieve a different final IP having a larger or smaller value of the criterion function. Having started with different initial IP's in a couple of cases, final IP's were reached which, although they were slightly different in composition, have nearly the same values for the criterion function.

The algorithm is relatively economical in terms of CPU time, requiring 2-3 minutes for a plan with eighteen receivers. The major factor in determining the computer time used in selecting the plan is not the number of receivers in the plan but the number of receivers which must be considered as possibly being added to the plan for improvement. The output of the program implementing the algorithm is a set of M receiver sites, the value of the criterion function, the differential contribution to the criterion function for each receiver at each trajectory point, the trajectory intervals during which each receiver should produce usable data, and a computer plot of the receiver and transmitter sites along with the ground track of the nominal flight path.

2. DOVAP MEASUREMENT MODEL. The DOVAP measurement system is a two way CW doppler system which measures a loop range change or equivalently an average loop range rate of a target between consecutive sampling times t_i and t_{i+1}. The normal sampling interval is 50 msec. A DOVAP instrumentation path consists of a transmitter station and usually ten or more receiver stations. In addition the target being tracked carries a frequency doubling transponder. The DOVAP receiver station receives both the reference frequency from the transmitter and the frequency doubled signal from the target transponder. A frequency comparison of the two signals and the addition of a $5KH_Z$ bias signal is performed at the

receiver. The resulting signal is converted to digital form by counting the doppler cycles over a 50 msec time interval. The resulting biased doppler measurement is

$$m(t_{i+1}) = 14.72194\ s(t_i,t_{i+1}) + 50278 + N_{i+1} \tag{1}$$

where $s(t_i, t_{i+1})$ is the refraction distorted change in loop range over the 50 msec interval. The loop range is defined as the sum of the range from transmitter to target and the range from target to receiver. N_{i+1} is a measurement error which is assumed to be random with zero mean. Thus, except for refraction, for which a correction is applied, DOVAP measurements are considered to be unbiased. Subtracting the bias 50278 and dividing by 14.72194 yields the modified DOVAP measurement $m'(t_{i+1})$

$$m'(t_{i+1}) = s(t_i,t_{i+1}) + N_{i+1} \tag{2}$$

After correcting for refraction we have the measurement model for the α^{th} receiver.

$$z_\alpha(t_i,t_{i+1}) = g_\alpha(x_i,x_{i+1}) + v_\alpha(i+1) \tag{3}$$

where $g(x_i,x_{i+1})$ is the loop range change from t_i to t_{i+1} and the arguments x_i and x_{i+1} indicate the dependence on the position coordinates at these times. x_i is the position vector of the target in the launcher coordinate system.

The loop range change is modelled as follows. Let (x_T,y_T,z_T) be the coordinates of the electrical center of the ground transmitter antenna in the launcher coordinate system. Also, let (x_R,y_R,z_R) be the coordinates at the electrical center of the ground receiver antenna. The z coordinates of the surveyed transmitter and receiver sites are modified to obtain z_T and z_R. The amount of modification depends on the type of antenna used. let (x,y,z) be the launcher coordinates of the target transmitting and receiving antennas. Let R_T be the range from ground transmitter antenna to target receiving antenna and R_R be the range from target transmitting antenna to ground receiver antenna. R_R and R_T are given by

$$R_T(t_i) = [(x(t_i)-x_T)^2+(y(t_i)-y_T)^2+(z(t_i)-z_T)^2]^{\frac{1}{2}} \tag{4}$$

$$R_R(t_i) = [(x(t_i)-x_R)^2+(y(t_i)-y_R)^2+(z(t_i)-z_R)^2]^{\frac{1}{2}} \tag{5}$$

Then the loop range change $g(x_i,x_{i+1})$ is

$$g(x_i,x_{i+1}) = R_T(t_{i+1}) + R_R(t_{i+1}) - R_T(t_i) - R_R(t_i) \tag{6}$$

The partial derivatives required for processing DOVAP observations using the above measurement model are

$$(7)\qquad \begin{bmatrix} \frac{\partial g}{\partial x(t_i)} \\ \frac{\partial g}{\partial y(t_i)} \\ \frac{\partial g}{\partial z(t_i)} \end{bmatrix} = \begin{bmatrix} \frac{-(x(t_i)-x_T)}{R_T(t_i)} - \frac{(x(t_i)-x_R)}{R_R(t_i)} \\ \frac{-(y(t_i)-y_T)}{R_T(t_i)} - \frac{(y(t_i)-y_R)}{R_R(t_i)} \\ \frac{-(z(t_i)-z_T)}{R_T(t_i)} - \frac{(z(t_i)-z_R)}{R_R(t_i)} \end{bmatrix}$$

$$(8)\qquad \begin{bmatrix} \frac{\partial g}{\partial x(t_{i+1})} \\ \frac{\partial g}{\partial y(t_{i+1})} \\ \frac{\partial g}{\partial z(t_{i+1})} \end{bmatrix} = \begin{bmatrix} \frac{(x(t_{i+1})-x_T)}{R_T(t_{i+1})} + \frac{(x(t_{i+1})-x_R)}{R_R(t_{i+1})} \\ \frac{(y(t_{i+1})-y_T)}{R_T(t_{i+1})} + \frac{(y(t_{i+1})-y_R)}{R_R(t_{i+1})} \\ \frac{(z(t_{i+1})-z_T)}{R_T(t_{i+1})} + \frac{(z(t_{i+1})-z_R)}{R_R(t_{i+1})} \end{bmatrix}$$

4. RECEIVER EFFECTIVENESS. The covariance matrix A^{-1} of the DOVAP batch processor solution [2] may be written in the elementary form

$$(9)\qquad A^{-1} = \left(\sum_{i=2}^{N} \sum_{\alpha=1}^{M} J_\alpha(x_{i-1},x_i)\, J_\alpha^T(x_{i-1},x_i) \right)^{-1}$$

where $J_\alpha(x_{i-1},x_i)$ is the 3N vector

(10)
$$J_\alpha(x_{i-1},x_i) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ G_{\alpha 1}^T(x_{i-1}) \\ G_{\alpha 2}^T(x_i) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ (i-1) \text{ place} \\ i \text{ place} \\ \\ \\ \\ \end{matrix}$$

if the α^{th} receiver is used in the i^{th} interval and $J_\alpha(x_{i-1},x_i)=0$ if the α^{th} receiver is not used in the i^{th} interval. $G_{\alpha 1}^T(x_{i-1})$ is the partial derivative of the DOVAP measurement function $g_\alpha(x_{i-1},x_i)$ with respect to its first argument and $G_{\alpha 2}^T(x_i)$ is the partial derivative with respect to the second argument. A receiver may be used in the i^{th} interval only if its tracking elevation $\theta_\alpha(i)$ at the midpoint of the interval satisfies the condition $\theta_1 \leq \theta_\alpha(i) \leq \theta_2$. θ_1 and θ_2 are input to the program with θ_1 normally in the range 10°-20° and θ_2 normally in the range 70°-80°. The effect of adding or deleting a DOVAP receiver from the k^{th} interval is easily written using the vectors $J_\alpha(x_{i-1},x_i)$. Let A_+^{-1} be the covariance matrix after adding a receiver β to the k^{th} and let A_-^{-1} be the covariance after deleting a receiver from the k^{th} interval. A_+^{-1} and A_-^{-1} are given by

(11)
$$A_+^{-1} = A^{-1} - \frac{A^{-1}J_\beta(x_{k-1},x_k)\, J_\beta^T(x_{k-1},x_k)\, A^{-1}}{1 + J_\beta^T(x_{k-1},x_k)\, A^{-1} J_\beta(x_{k-1},x_k)}$$

(12)
$$A_-^{-1} = A^{-1} + \frac{A^{-1}J_\beta(x_{k-1},x_k)\, J_\beta^T(x_{k-1},x_k)\, A^{-1}}{1 - J_\beta^T(x_{k-1},x_k)\, A^{-1} J_\beta(x_{k-1},x_k)}$$

Thus the effect of adding or deleting a receiver β in the k^{th} interval is

(13) $$E(k,\beta) = \mp \frac{V(k,\beta)\ V^T(k,\beta)}{1 \pm q}$$

where the $V(k,\beta)$ is a 3N vector. The upper sign indicates addition of a receiver and the lower sign deletion of a receiver.

(14) $$V(k,\beta) = A^{-1} J_\beta(x_{k-1}, x_k)$$

and q is the scalar

(15) $$q = J_\beta^T(x_{k-1}, x_k)\ A^{-1} J_\beta(x_{k-1}, x_k)$$

Partition $V(k,\beta)$ into its 3×1 subvectors

(16)
$$V(k,\beta) = \begin{bmatrix} V_1(k,\beta) \\ V_2(k,\beta) \\ \vdots \\ V_N(k,\beta) \end{bmatrix}$$

The 3-vectors $V_j(k,\beta)$ are easily computed as

(17) $$V_j(k,\beta) = A^{-1}_{j,k-1} G^T_{\beta 1}(x_{k-1}) + A^{-1}_{j,k} G^T_{\beta 2}(x_k)$$

where A^{-1}_{ij} are 3×3 block elements of A^{-1}. The scalar q is given by

(18) $$q = G_{\beta 1}(x_{k-1})\ V_{k-1}(k,\beta) + G_{\beta 2}(x_k)\ V_k(k,\beta)$$

The effect matrix $E(k,\beta)$ can be written in terms 3×3 block elements $E_{ij}(k,\beta)$

(19) $$E_{ij}(k,\beta) = \mp \frac{V_i(k,\beta)\ V_j^T(k,\beta)}{1 \pm q}$$

Define the vector

$$V_i^*(k,\beta) = \frac{V_i(,\beta)}{\sqrt{1 \pm q}} \tag{20}$$

Then

$$E_{ij}(k,\beta) = \mp V_i(k,\beta)\ V_j^{*T}(k,\beta) \tag{21}$$

This expression for the effect of adding or deleting the β^{th} receiver in the k^{th} interval is used to modify the covariance matrix A^{-1}. The above equations are used sequentially to add or delete the β^{th} receiver for all intervals k=2,N. In adding or deleting in the $(k+1)^{st}$ interval, A is the modified covariance obtained in adding or deleting from the k^{th} interval.

The effectiveness of a receiver is defined as

$$e(\beta) = \left| \sum_{k=2}^{N} \mathrm{Tr} E(k,\beta) \right| \tag{22}$$

Using (13), (16), and (20) in (22) $e(\beta)$ can be written as

$$e(\beta) = \sum_{k=2}^{N} \sum_{i=1}^{N} \mathrm{Tr}\ V_i^*(k,\beta)\ V_i^{*T}(k,\beta) \tag{23}$$

or

$$e(\beta) = \sum_{k=2}^{N} \sum_{i=1}^{N} V_i^{*T}(k,\beta)\ V_i^*(k,\beta) \tag{24}$$

More generally, the receiver effectiveness may be defined giving more importance to some trajectory points than others by using weights $W_i>0$. Then we define the receiver effectiveness as

$$e(\beta) = \sum_{k=2}^{N} \sum_{i=1}^{N} W_i V_i^{*T}(k,\beta)\ V_i^*(k,\beta) \tag{25}$$

Another quantity which is useful in analyzing the effects of a receiver in a DOVAP IP is the differential effect of a receiver which is defined as the increase in the diagonal elements of A^{-1} if the receiver is deleted from the IP. Thus the differential effect is the diagonal elements of

$$E(\beta) = \sum_{k=2}^{N} E(k,\beta) \tag{26}$$

Also of interest are the quantities

$$e_j(k,\beta) = \sum_{i=1}^{N} V_{ij}^{*2}(k,\beta) \qquad j=1,2,3 \tag{26}$$

the numbers $e_j(k,\beta)$ for a given interval k show the importance of the observations from receiver β during that interval.

5. OPTIMAL RECEIVER SELECTION. The optimal receiver selection procedure chooses a set of M existing receiver sites which minimizes

$$c = \sum_{i=1}^{N} w_i \operatorname{tr} \operatorname{cov}(x_i) \tag{27}$$

The quantity c may be interpreted as a weighted sum of the error estimates which would be obtained if loop range change observations of a nominal flight path from a set of m receivers were processed with the DOVAP batch processor. The choice of m(10-20) receivers from the 650 receivers presently available at WSMR is a very large combinatorial programming problem which is prohibitive from a computational point of view. Even if the number of available receivers is considerably reduced by some intelligent screening procedures, the magnitude of the problem is still prohibitive. Rather than trying to select the set of receivers which achieves the global minimum of c we will use an IP improvement algorithm to obtain a local minimum of c.

The IP improvement algorithm selects receivers from an instrumentation planning pool (IPP). The IPP is a set of receivers which is obtained from the set of all existing sites by placing constraints on the tracking elevation of a receiver and on the transmitter reference signal strength available at a receiver site. The following constraints are used to develop the IPP. Let $T_\alpha=1$ if the α^{th} receiver can receive a sufficiently strong 36.2 MH_z reference signal from the transmitter. Otherwise, $T_\alpha=0$. The values of T_α for all receivers and transmitters were furnished by the DOVAP instrumentation system personnel from tests conducted to determine signal strength. The 72.4 MH_z antennas used at the DOVAP receiver sites may have nulls at low and/or high elevation angles of the incoming 72.4 MH_z signal. Thus, in order to insure a sufficient 72.4 MH_z signal strength at the receiver, the elevation angle of the line of sight between the receiver and the trajectory should be between θ_1 and θ_2. The minimum θ_1 and the maximum angle θ_2 are a function of range between receiver and target. For each receiver α let RG_α be the shortest ground range for any of the points on the nominal trajectory.

$$R_{G_\alpha} = \min_{i=1,N} \left[(x_i - x_{R_\alpha})^2 + (y_i - y_{R_\alpha})^2\right]^{\frac{1}{2}} \tag{28}$$

where (x_i, y_i, z_i) i=1,N are the coordinates of the trajectory points and $(x_{R_\alpha}, y_{R_\alpha}, z_{R_\alpha})$ are the coordinates of the α^{th} receiver. Let E_α be the elevation of the line of sight from α^{th} receiver to the trajectory point corresponding to R_{G_α}.

$$E_\alpha = \tan^{-1} \frac{(z_i - z_{R_\alpha})}{R_{G_\alpha}} \tag{29}$$

The IPP is given by

$$IPP = \{\alpha | T_\alpha = 1, \theta_1 \leq E_\alpha \leq \theta_2\} \tag{30}$$

Let M be the maximum number of DOVAP receiver sites we wish to select and let $M_1 \leq M$ be the number of receivers used to start the selection procedure. M_1 may be considered as the minimum number of receivers to be selected. Given an arbitrary initial IP having M_1 receivers the SELECT program will construct an IP having M_1+1 receivers by adding the receiver from those available in the IPP which results in the greatest decrease in

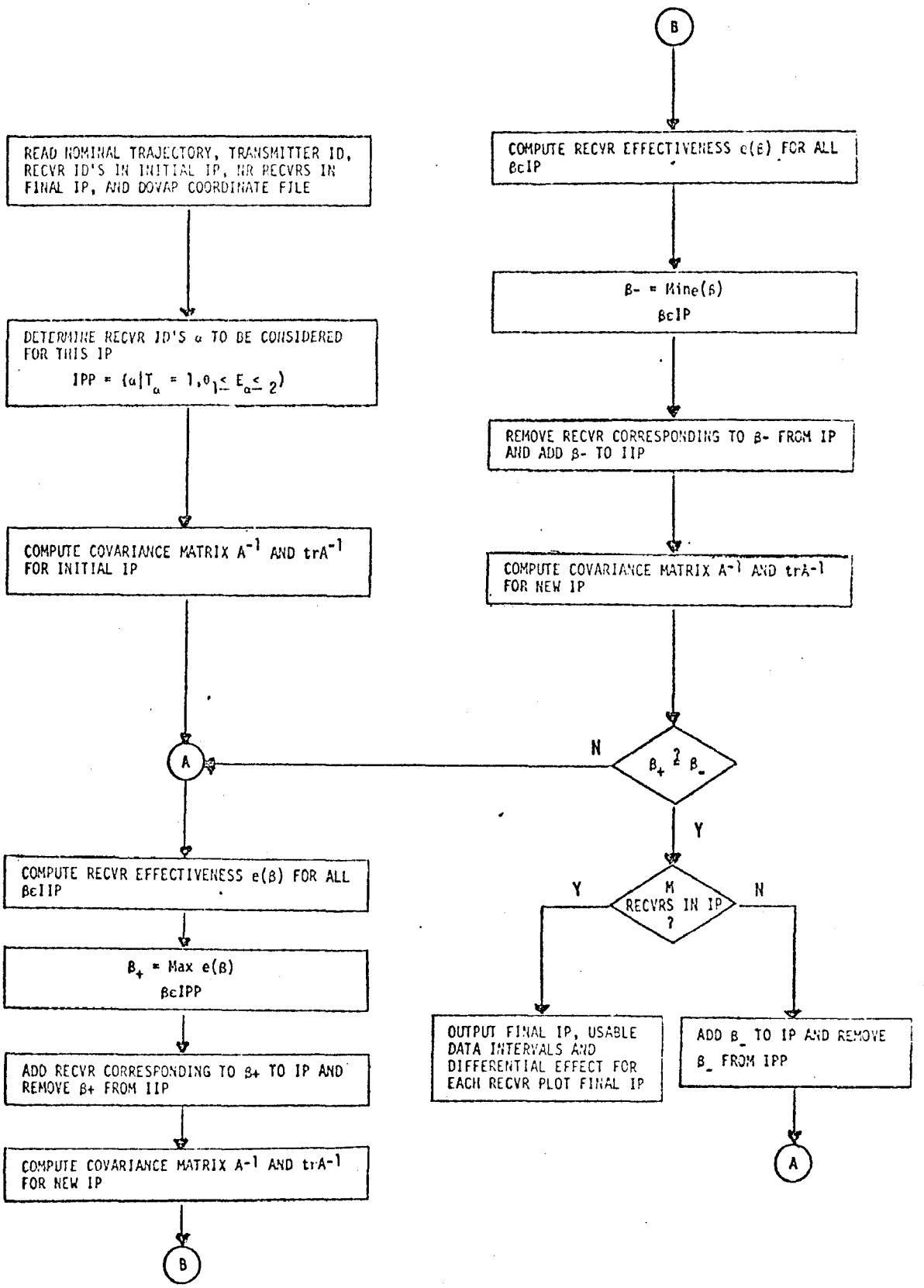
$$c = \sum_{i=1}^{N} w_i \, tr \, cov(\bar{x}_i)$$

The SELECT program then deletes the receiver from this modified IP which results in the smallest increase in c. This exchange procedure between the IPP and the IP is continued until no further improvement can be made. The exchange procedure terminates when the received added to the IP is identical to the receiver deleted from the IP. The final set of receivers in the IP is the best set of M_1 receivers. To form the best set of M_1+1 receivers start with the IP formed by adding the receiver added and deleted at the termination of the previous stage to the best set of M_1 receivers. Proceed with the exchange process. Similarly, the best IP's having M_1+2, M_1+3, $\cdots$, M receivers are obtained. The following flow chart may clarify the above description of the instrumentation plan improvement algorithm.

6. REFERENCES.

a. Agee, W. S. and R. H. Turner, "DOVAP Best Estimate of Trajectory", Technical Report Number 55, WSMR, May 75.

b. Agee, W. S. and J. L. Meyer, "DOVAP Instrumentation Planning", Technical Report Number 56, WSMR, May 75.


READ NOMINAL TRAJECTORY, TRANSMITTER ID, RECVR ID'S IN INITIAL IP, NR RECVRS IN FINAL IP, AND DOVAP COORDINATE FILE
DETERMINE RECVR ID'S α TO BE CONSIDERED FOR THIS IP
IPP = (α|T_α = 1, θ_1 ≤ E_α ≤ θ_2)
COMPUTE COVARIANCE MATRIX A^-1 AND trA^-1 FOR INITIAL IP
A
COMPUTE RECVR EFFECTIVENESS e(β) FOR ALL β∈IIP
β+ = Max e(β) β∈IPP
ADD RECVR CORRESPONDING TO β+ TO IP AND REMOVE β+ FROM IIP
COMPUTE COVARIANCE MATRIX A-1 AND trA-1 FOR NEW IP
B
B
COMPUTE RECVR EFFECTIVENESS e(β) FOR ALL β∈IP
β- = Mine(β) β∈IP
REMOVE RECVR CORRESPONDING TO β- FROM IP AND ADD β- TO IIP
COMPUTE COVARIANCE MATRIX A^-1 AND trA-1 FOR NEW IP
β+ ≥ β-
N
Y
M RECVRS IN IP ?
Y
N
OUTPUT FINAL IP, USABLE DATA INTERVALS AND DIFFERENTIAL EFFECT FOR EACH RECVR PLOT FINAL IP
ADD β- TO IP AND REMOVE β- FROM IPP
A

PROVING PROGRAMS CORRECT

ELWOOD D. BAAS

ARMY MISSILE TEST AND EVALUATION
WHITE SANDS MISSILE RANGE, NEW MEXICO 88002

ABSTRACT

Mathematical methods of proving programs correct have recently been investigated by several interdisciplinary groups. This effort has been motivated by the fact that some computer applications are being restricted in important areas because of the inability to design and implement software programs which can be shown convincingly to be correct. It is also recognized that debugging and maintaining computer programs are two very serious and costly problems facing the computer industry. This paper provides a survey of the investigative work completed to date and some new program design concepts which have resulted from the effort.

Proving programs correct (or program correctness) can rightly be considered as part of what is currently referred to as structured programming. The other areas related to structured programming which are discussed in the literature could be categorized as: programming methodology, program notation, and program verification. It is within this general context that we shall address the topic of program correctness.

Using 20/20 hindsight, it is obvious that during the first 25 years of programming there was too little emphasis placed on program correctness and too much emphasis on debugging. As the size and cost of large computer programs burgeoned during the past 5-10 years, more and more effort was expended in trying to develop a degree of structural integrity along with a given computer program. In most cases this search for a viable structure was motivated by a very practical reason: program debugging was taking so long that the program or system became obsolete before the program was checked out!

At present there are over one hundred people (worldwide), mainly in the areas of Mathematics and Computer Science, who are actively engaged in the areas of proving assertions about programs. Some of the primary methods of formal program proof construction currently under investigation are:

(1) Inductive Methods
(2) Calculus Schemata
(3) Graph Techniques
(4) Recursive Schemes
(5) Radix Sorting Techniques
(6) Algebraic Models

These methods involve the translation of the program to higher level semantics which then can be used to verify the coding logic independent of specific inputs and outputs. Thus one is concerned not so much at how a program changes values of variables, but instead at how relations

amoung variables remain the same. At the higher level construct, precise definitions of completeness, consistency or correctness can be applied to draw conclusions concerning the original program and its input-output relations. Strong and Walker of IBM summarize this basic notion as follows: "Most approaches to proving programs correct concentrate on the verification of input-output relations, the verification techniques being independent of any particular attributes of the input-output relation in question."[1]

One should note at this point that the concept of formally proving a program correct is radically different than the usual process of testing or debugging a program. Testing can prove that a program is incorrect, but no reasonable amount of testing can ever prove that a non-trivial program will be correct for all allowable variations of input conditions. Professor Dijkstra states the case clearly: "Program testing can be a very effective way to show the presence of bugs, but is is hopelessly inadequate for showing their absence."[2]

There are some serious limitations to these formal correctness methods as far as practical applications are concerned. First, there is no universal higher level scheme at present; each construct is dependent on the particular program language used. Thus each correctness scheme must be custom built around a given language. In fact, Ashcroft, has shown that the limitations and possibilities of correctness methods are just those of language definition methods. As a practical example of a limitation, consider the frequent question concerning program proofs: "What about overflow?" The answer is either, "this particular method is defining a simple language with idealized storage registers" or "show me a language definition that considers overflow and I will construct a correctness method which takes care of it."[3]

A second limitation of proving assertions about programs has to do with program size. As the programs become larger, the proofs become lengthy. When the techniques of proof become long and detailed, they

may themselves contain errors. Jones comments on this problem:

APPENDIX

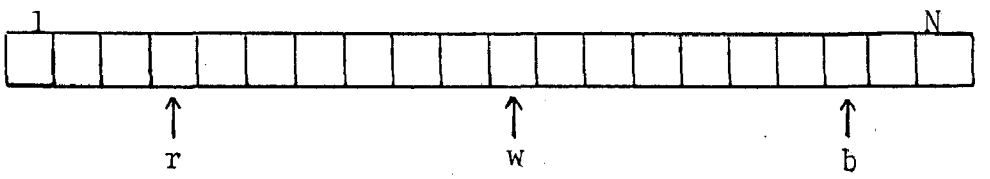
Dutch National Flag Problem

(This programming problem and solution were presented by Professor Dijkstra at the Conference on Programming Methodology held at the University of New Mexico, Albuquerque, NM, Jan 7-11, 1974.)

Problem: Each of N buckets contains a single red, white, or blue pebble. Using a function swap (p,q) which permutes pebbles, and a function look (p) which identifies the color of a pebble, write a program which will rearrange the pebbles so that red, white, and blue are separated into three bins.

Restrictions: Each pebble may be looked at only once. Missing colors are allowed. No arrays are allowed.

Solution:



Initial conditions: r: = 1; w: = N, b: = N

```
Program: do  w ≥ r ⟶
             col: = look (w);
               if  col = red ⟶ swap (r,w); inc(r)[]
                   col = white ⟶ dec (w)[]
                   col = blue ⟶ swap (w,b); dec (w,b)
               fi
         od
```

Note: This program is machine independent. To prove this program correct one needs to simply check that the "or statements" are exclusive and that the program terminates.

REFERENCES

1. Strong, H.R., Walker, S.A., "Properties Preserved Under Recursion Removal," Proceedings of an ACM Conference on Proving Assertions about Programs, P. 97, 1972.

2. Dijkstra, E., "The Humble Programmer," Communications of the ACM, Vol 15, No. 10, October 1972.

3. Ashcroft, E.A., "Program Correctness Methods and Language Definition," Proceedings of an ACM Conference on Proving Assertions about Programs, P. 51-58, 1972.

4. Jones, C.B., "Formal Development of Correct Algorithms," Proceedings of an ACM Conference on Proving Assertions about Programs, P. 150, 1972.

5. Gorman, T.P., "A Software Engineering Approach to the Space Information System of the Future," Software Engineering, Academic Press, N.Y., 1970.

6. Dijkstra, E., "The Humble Programmer," Communications of the ACM, Vol 15, No. 10, October 1972.

7. Ibid

8. Dijkstra, E., EWD 249 - Notes on Structured Programming, T.H. Report 70, Technological University, Eindhoven Netherlands, 1970.

9. Mills, H.D., Mathematical Foundations for Structured Programming, Federal Systems Division, IBM, Gaithersburg, Maryland, 1972.

10. Linden, T.A., "A Summary of Progress Toward Proving Program Correctness," Fall Joint Computer Conference, Anaheim, California, December 1972.

11. Bradshaw, F.T., Some Structural Ideas for Computer Systems, Case Western Reserve University, Cleveland, Ohio, 1972.

GENERALIZED PLANE STRAIN IN AN ELASTIC, PERFECTLY PLASTIC CYLINDER, WITH REFERENCE TO THE HYDRAULIC AUTOFRETTAGE PROCESS

Alexander S. Elder
Interior Ballistics Laboratory

Robert C. Tompkins
Thomas L. Mann
Applied Mathematics and Sciences Laboratory
Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland

ABSTRACT

Conditions for generalized plane strain in an elastic, perfectly plastic cylinder subject to uniform internal pressure are derived from specific assumptions concerning the displacements and the end conditions. Specifically, we assume the tangential displacement is a function of the radius only, and the axial displacement is a function of the distance from the diametral plane of reference. We conclude immediately the shear strains and rotations are zero. The shear stresses are also zero if the cylinder is free of residual stresses in its original state. Symmetry of the normal stresses follow from arguments involving the Prandtl-Reuss flow equations, the Von Mises yield condition, and the elastic behavior of the plastic zone when subject to hydrostatic pressure. The equilibrium equations, boundary conditions, and end conditions are also involved. The axial strain is found to be independent of r and z. Three equations governing plastic flow are derived which are similar in form to those given by Prager and Hodge for the plane strain condition. These equations were solved numerically by integration along the characteristics. An iterative procedure was required to determine the radial pressure and axial strain at the elastic plastic interface in a manner which satisfied both boundary and end conditions. Numerical results are presented for a cylinder with a wall ratio of two for both open end and closed end conditions. The residual stresses which remain after release of the hydraulic pressure were also calculated. Axial as well as circumferential residual stresses are produced. Calculations for wall ratios in the range 1.5-3.0 were carried out. Re-yielding at the inner surface will occur if the wall ratio exceeds 2.25; the stresses remain elastic for smaller wall ratios when the pressure is released.

I. INTRODUCTION

In this paper we consider axial, circumferential, and radial stresses in an elastic, perfectly plastic cylinder pressurized internally until the entire cylinder undergoes plastic yielding. It is assumed the cylinder is in a condition of generalized plane strain, the ends of the cylinder are either open or closed, and appropriate boundary conditions apply at the cylindrical surfaces. The mathematical treatment is a generalization of the plane strain analysis of Prager and Hodge. The numerical analysis is based on integration along the characteristics of a system of quasi-linear, hyperbolic, partial differential equations. The residual stresses which exist when the hydraulic pressure is released are also calculated.

The analysis begins with three assumptions concerning the displacement field in a long, uniformly pressurized cylinder originally free of residual stresses, and deduces the nature of the stress and strain fields which are consistent with these assumptions. As usual, boundary conditions, equilibrium and compatibility equations, and equations giving the strains as functions of the stresses or displacements lead to the required stress and strain fields in the elastic case. Initial conditions, the yield condition, and equations governing plastic flow are essential additional requirements for the plastic zone. The solution obtained from these relations, together with the original assumptions concerning the displacements, yields a mathematical solution which is internally consistent. Conditions of confined plasticity are assumed; that is, the plastic strains are of the same order of magnitude as the elastic strains, and can be derived from the displacements in the same manner. For simplicity, infinitesimal analysis is used throughout, perhaps at some cost in realism when one considers the significant elastic strains which can occur in a pressurized cylinder of very high strength steel.

II. GENERALIZED PLANE STRAIN IN AN ELASTIC CYLINDER

In an elastic, isotropic cylinder, we assume: the tangential displacement is zero, the radial displacement is a function of the radius only, and the axial displacement is a function of the distance from the diametral plane of reference.

$$v = 0 \tag{1}$$

$$u = f(r) \tag{2}$$

$$w = g(z) , \tag{3}$$

The displacement-strain relations, equations of equilibrium, and stress-strain relations are used to show that the stresses and strains are axially symmetric and torsion-free.

The shear strains are given by[1]

$$\gamma_{r\theta} = \frac{\partial u}{r\partial\theta} + \frac{\partial v}{\partial r} - \frac{v}{r} \tag{4}$$

$$\gamma_{rz} = \frac{\partial u}{\partial z} + \frac{\partial w}{\partial r} \tag{5}$$

$$\gamma_{z\theta} = \frac{\partial v}{\partial z} + \frac{1}{r}\frac{\partial w}{\partial \theta} \, . \tag{6}$$

On referring to Eqs. (1) - (3), we see

$$\frac{\partial v}{\partial r} = 0, \quad \frac{\partial v}{\partial \theta} = 0, \quad \frac{\partial v}{\partial z} = 0 \tag{7}$$

$$\frac{\partial u}{\partial \theta} = 0, \quad \frac{\partial u}{\partial z} = 0 \tag{8}$$

$$\frac{\partial w}{\partial \theta} = 0, \quad \frac{\partial w}{\partial r} = 0 \tag{9}$$

so that

$$\gamma_{r\theta} = 0, \quad \gamma_{rz} = 0, \quad \gamma_{z\theta} = 0. \tag{10}$$

The stress-strain relations for shear are

$$\tau_{r\theta} = G\gamma_{r\theta}, \quad \tau_{rz} = G\gamma_{rz}, \quad \tau_{z\theta} = G\gamma_{z\theta} \tag{11}$$

so that

$$\tau_{r\theta} = 0, \quad \tau_{rz} = 0, \quad \tau_{z\theta} = 0 \, . \tag{12}$$

Eqs. (7) thru (12) show that the cylinder does not undergo shearing strains due to torsion.

Next, we use the stress-strain laws to show the stresses and strains are axially symmetric. The normal strains are given by

$$\varepsilon_r = \frac{\partial u}{\partial r} \, , \quad \varepsilon_\theta = \frac{u}{r} \, , \quad \varepsilon_z = \frac{\partial w}{\partial z} \neq 0 \tag{13}$$

since $\frac{\partial v}{\partial \theta} = 0$. The stress-strain equations for the normal stresses are given by

---

1. *Timoshenko, S. and Goodier, I.N., Theory of Elasticity, second edition, McGraw Hill Book Company, Inc., New York, 1951. Pages 305, 306.*

$$\sigma_r = K(\varepsilon_\theta+\varepsilon_r+\varepsilon_z) + \frac{2}{3} G(2\varepsilon_r-\varepsilon_\theta-\varepsilon_z) \tag{14}$$

$$\sigma_\theta = K(\varepsilon_\theta+\varepsilon_r+\varepsilon_z) + \frac{2}{3} G(2\varepsilon_\theta-\varepsilon_z-\varepsilon_r) \tag{15}$$

$$\sigma_z = K(\varepsilon_\theta+\varepsilon_r+\varepsilon_z) + \frac{2}{3} G(2\varepsilon_z-\varepsilon_r-\varepsilon_\theta) . \tag{16}$$

On referring to Eqs. (7), (8), and (9), we see

$$\frac{\partial \varepsilon_r}{\partial \theta} = 0, \quad \frac{\partial \varepsilon_\theta}{\partial \theta} = 0, \quad \frac{\partial \varepsilon_z}{\partial \theta} = 0 \tag{17}$$

and consequently,

$$\frac{\partial \sigma_r}{\partial \theta} = 0, \quad \frac{\partial \sigma_\theta}{\partial \theta} = 0, \quad \frac{\partial \sigma_z}{\partial \theta} = 0 . \tag{18}$$

The normal strains and stresses, as well as the displacements, are axially symmetric. We also find that

$$\frac{\partial \varepsilon_\theta}{\partial z} = 0, \quad \frac{\partial \varepsilon_r}{\partial z} = 0, \quad \frac{\partial \varepsilon_z}{\partial r} = 0 . \tag{19}$$

The equations of equilibrium are required in order to complete the solution. We have

$$\begin{aligned}
&\frac{\partial \sigma_r}{\partial r} + \frac{1}{r}\frac{\partial \tau_{r\theta}}{\partial \theta} + \frac{\partial \tau_{rz}}{\partial z} + \frac{\sigma_r - \sigma_\theta}{r} = 0 \\
&\frac{\partial \tau_{rz}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta z}}{\partial \theta} + \frac{\partial \sigma_z}{\partial z} + \frac{\tau_{rz}}{r} = 0 \\
&\frac{\partial \tau_{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_\theta}{\partial \theta} + \frac{\partial \tau_{\theta z}}{\partial z} + \frac{2\tau_{r\theta}}{r} = 0
\end{aligned} \tag{20}$$

which reduce to

$$\frac{\partial \sigma_r}{\partial r} + \frac{\sigma_r - \sigma_\theta}{r} = 0 \tag{21}$$

$$\frac{\partial \sigma_z}{\partial z} = 0 \tag{22}$$

$$\frac{\partial \sigma_\theta}{\partial \theta} = 0 . \tag{23}$$

We see that Eq. (23) is redundant, as it duplicates information given in Eq. (18).

From Eq. (16) and Eq. (22) we find

$$\frac{\partial \varepsilon_z}{\partial z} = 0 \tag{24}$$

so that

$$w = C + Dz \ . \tag{25}$$

An equation for the radial displacement is found from Eqs. (13), (14), (15) and (21).

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} - \frac{u}{r^2} = 0 \tag{26}$$

$$u = Ar + B/r \tag{27}$$

$$\varepsilon_\theta = A + B/r^2 \tag{28}$$

$$\varepsilon_r = A - B/r^2 \tag{29}$$

and from Eq. (25)

$$\varepsilon_z = D \ . \tag{30}$$

The Lame' formulas for the stresses are obtained from Eqs. (14), (15), and (16)

$$\sigma_r = A' + B'/r^2 \tag{31}$$

$$\sigma_\theta = A' - B'/r^2 \tag{32}$$

$$\sigma_z = D' \tag{33}$$

where

$$A' = K(2A+D) + \frac{2}{3}G(A-D) \tag{34}$$

$$B' = -2GB \tag{35}$$

$$D' = K(2A+D) - \frac{4}{3}G(A-D) . \tag{36}$$

The constants must be determined from the boundary conditions which may be stated either in terms of stresses or displacements. We also find

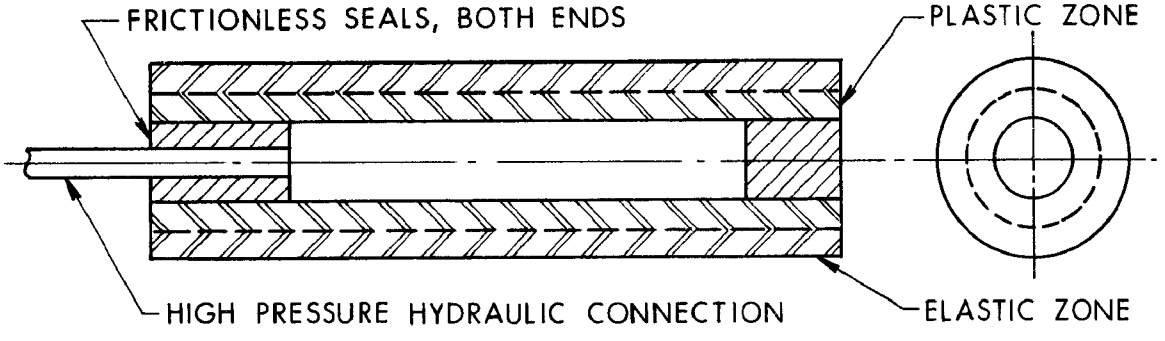
$$\frac{\partial \sigma_z}{\partial r} = 0, \qquad \frac{\partial(\sigma_r+\sigma_\theta)}{\partial r} = 0 , \tag{37}$$

The stress-strain relations given by Eqs. (14), (15) and (16) are not valid in an elastic-plastic cylinder, so a more detailed analysis is required to prove the stresses and strains are independent of z and θ. Eq. (37) is also invalid for a long cylinder when plastic flow occurs.

The existence of residual stresses resulting from prior thermal or mechanical history does not affect the preceding formulas provided the total stresses, due to residual stresses and due to loading, remain within the elastic limit. We assume the stresses, strains, and displacements given above are produced by external loads only.

III. GENERALIZED PLANE STRAIN IN AN ELASTIC-PERFECTLY PLASTIC CYLINDER

Consider a long, hollow cylinder pressurized internally until a plastic zone develops near the inner surface, as shown below in Figure 1. We neglect conditions near the ends where hydraulic seals and various mechanical attachments will produce a complicated two-dimensional stress distribution. A few diameters away from the ends, these local effects will disappear, and the assumed conditions governing the displacements, given by Eqs. (1), (2), and (3) of the preceding section, will closely approximate the actual physical conditions.



NOTE: EXTERNAL CYLINDRICAL JACKET NOT SHOWN.

FIGURE 1. Cylinder Deformed Plastically by Internal Hydraulic Pressure

Eqs. (7), (8), (9) and also (13) of the preceding section are valid, as they do not involve the stress-strain relations or the yield condition. Additional analysis is required to show that the shear stresses are zero.

The Prandtl-Reuss flow equation is assumed to govern the deformation in the shear during plastic flow[2]

$$G\dot{\gamma}_{rz} = \dot{\tau}_{rz} + \lambda\tau_{rz} \tag{39}*$$

$$G\dot{\gamma}_{z\theta} = \dot{\tau}_{z\theta} + \lambda\tau_{z\theta} \tag{40}$$

$$G\dot{\gamma}_{\theta r} = \dot{\tau}_{\theta r} + \lambda\tau_{\theta r} . \tag{41}$$

In these equations, λ is an unknown parameter depending on time and the coordinates r,z, and θ. Since the shear strains are zero, we have

$$\dot{\tau}_{rz} + \lambda\tau_{rz} = 0 \tag{42}$$

$$\dot{\tau}_{z\theta} + \lambda\tau_{z\theta} = 0 \tag{43}$$

$$\dot{\tau}_{\theta r} + \lambda\tau_{\theta r} = 0 . \tag{44}$$

On solving we find

$$\tau_{rz} = \alpha_{rz}e^{-\delta}, \quad \tau_{z\theta} = \alpha_{z\theta}e^{-\delta}, \quad \tau_{\theta r} = \alpha_{\theta r}e^{-\delta} \tag{45}$$

where

$$\delta = \int_0^t \lambda dt . \tag{46}$$

We assume the cylinder is free of residual stresses at the beginning of pressurization. Then the shear stresses will be zero while the tube remains elastic and during incipient plastic flow. If we measure time from the beginning of plastic flow, we have

$$\tau_{rz} = 0, \quad \tau_{z\theta} = 0, \quad \tau_{\theta r} = 0 \quad \text{when } t=0. \tag{47}$$

---

2. *Prager, W. and Hodge, Jr., P.G., Theory of Perfectly Plastic Solids, Dover Publications, New York, 1968. Pages 16-32, 95-122.*

* *In this section the superimposed dot means* $\frac{d}{dt}$.

Now λ is a positive scalar, so $\lambda>0$ when $t>0$. Hence, the exponential factor $e^{-\delta}$ in Eq. (45) cannot vanish. It follows that

$$\alpha_{rz} = 0, \quad \alpha_{z\theta} = 0, \quad \alpha_{\theta r} = 0 \tag{48}$$

and consequently

$$\tau_{rz} = 0, \quad \tau_{z\theta} = 0, \quad \tau_{\theta r} = 0 \text{ for } t\geq 0 . \tag{49}$$

Eq. (49) corresponds to Eq. (12) of the first section. Eqs. (21), (22) and (23) are also valid since the shear stresses are zero; these equations are repeated below for convenience.

$$\frac{\partial \sigma_r}{\partial r} + \frac{\sigma_r - \sigma_\theta}{r} = 0$$

$$\frac{\partial \sigma_z}{\partial z} = 0$$

$$\frac{\partial \sigma_\theta}{\partial \theta} = 0.$$

As mentioned previously, we cannot use the elastic stress-strain laws to prove the stresses and strains are independent of z and θ. However, if we add Eqs. (14), (15) and (16), we find

$$\sigma_r + \sigma_\theta + \sigma_z = 3K(\varepsilon_\theta + \varepsilon_r + \varepsilon_z). \tag{50}$$

This relation between the sum of the normal stresses and the sum of the normal strain is valid in both the elastic and plastic zones, as it does not involve the deviatoric stresses and strains.

The first equilibrium equation gives

$$\sigma_\theta = r \frac{\partial \sigma_r}{\partial r} + \sigma_r \tag{51}$$

and from the displacement strain we have

$$\varepsilon_\theta = \frac{u}{r}, \quad \varepsilon_r = \frac{\partial u}{\partial r}$$

so that

$$r\left(r \frac{\partial \sigma_r}{\partial r} + 2\sigma_r\right) + r\sigma_z = 3K\left(r \frac{\partial u}{\partial r} + u + r \frac{\partial w}{\partial z}\right). \tag{52}$$

On integrating with respect to r, we find

$$r^2\sigma_r + \int r\sigma_z dr = 3K(ru + \int r \frac{\partial w}{\partial z} dr). \tag{53}$$

Integration by parts gives

$$\int r \frac{\partial w}{\partial z} dr = \frac{1}{2} r^2 \frac{\partial w}{\partial z} - \frac{1}{2} \int r^2 \frac{\partial^2 w}{\partial r \partial z} dr. \tag{54}$$

The last integral vanishes since we assume

$$\frac{\partial w}{\partial r} = 0,$$

Hence,

$$r^2\sigma_r + \int r\sigma_z dr = 3K(ru + \frac{1}{2} r^2 \frac{\partial w}{\partial z}), \tag{55}$$

From the boundary conditions

$$\sigma_r = - p_i \quad \text{at } r = a$$

$$\sigma_r = 0 \quad \text{at } r = b$$

which is valid for the part of the cylinder between the seals, we find

$$p_i a^2 + \int_a^b r\sigma_z dr = 3K(ru)\Big|_a^b + \frac{1}{2} (b^2 - a^2) \frac{\partial w}{\partial z}, \tag{56}$$

Evidently $p_i a^2$ is independent of z. The integral is proportional to the net axial load applied at the ends of the cylinder, and must be independent of z from equilibrium conditions. The term on the right hand of the equation involving u is also independent of z. Hence,

$$\frac{\partial w}{\partial z} = D$$

and the axial strain is independent of z. The conditions

$$\frac{\partial \varepsilon_z}{\partial z} = 0, \quad \frac{\partial \varepsilon_z}{\partial r} = 0$$

define the conditions for generalized plane strain.

On returning to Eq. (50), we have

$$\sigma_r + \sigma_\theta + \sigma_z = 3K\left(\frac{u}{r} + \frac{\partial u}{\partial r} + \varepsilon_z\right)$$

so that

$$\frac{\partial(\sigma_r+\sigma_\theta)}{\partial z} = 0, \tag{57}$$

We use the Von Mises yield condition to prove that σ_r and σ_θ are independent of z. In the absence of shear stresses the Von Mises yield condition is given by

$$(\sigma_\theta-\sigma_r)^2 + (\sigma_r-\sigma_z)^2 + (\sigma_z-\sigma_\theta)^2 = 6k^2 \tag{58}$$

Eq. (58) is assumed to be valid for the entire plastic zone. In this equation, the normal stresses include residual stresses as well as stresses produced by the external forces on the cylinder. We must assume that there are no residual stresses when the cylinder is in its original, unloaded state; otherwise the use of the Lame' formulas to calculate stresses at the elastic-plastic interface, to be discussed later, would be invalid.

$$\sigma_r = 0, \quad \sigma_\theta = 0, \quad \sigma_z = 0 \tag{59}$$

when the cylinder is in its original elastic state and not subject to external loads.

Differentiate Eq. (58) with respect to z; we find

$$(2\sigma_\theta-\sigma_r-\sigma_z)\frac{\partial\sigma_\theta}{\partial z} + (2\sigma_r-\sigma_\theta-\sigma_z)\frac{\partial\sigma_r}{\partial z} = 0 \tag{60}$$

since $\frac{\partial\sigma_z}{\partial z} = 0$. On combining Eqs. (57) and (60), we find

$$(\sigma_r-\sigma_\theta)\frac{\partial\sigma_\theta}{\partial z} = 0, \quad (\sigma_\theta-\sigma_r)\frac{\partial\sigma_r}{\partial z} = 0. \tag{61}$$

We have two alternatives. If

$$\sigma_\theta \neq \sigma_r,$$

then

$$\frac{\partial\sigma_\theta}{\partial z} = 0, \quad \frac{\partial\sigma_r}{\partial z} = 0 \tag{62}$$

If on the other hand

$$\sigma_\theta = \sigma_r,$$

then Eq. (58) gives

$$\sigma_r - \sigma_z = \pm\sqrt{3}\,k, \quad \sigma_\theta - \sigma_z = \pm\sqrt{3}\,k \tag{63}$$

so that Eq. (62) is again satisfied since $\frac{\partial \sigma_z}{\partial z} = 0$. Hence, Eq. (62) is satisfied under all conditions.

A similar line of reasoning can be used to show the stresses are axially symmetric. We have from Eq. (50)

$$\frac{\partial(\sigma_z+\sigma_r)}{\partial\theta} = 0 \tag{64}$$

Eq. (58) gives

$$(2\sigma_z - \sigma_r - \sigma_\theta)\frac{\partial \sigma_z}{\partial\theta} + (2\sigma_r - \sigma_\theta - \sigma_z)\frac{\partial \sigma_r}{\partial\theta} = 0 \tag{65}$$

Eqs. (64) and (65) give

$$(\sigma_z - \sigma_r)\frac{\partial \sigma_z}{\partial\theta} = 0, \qquad (\sigma_r - \sigma_z)\frac{\partial \sigma_r}{\partial\theta} = 0 \tag{66}$$

and we finally obtain

$$\frac{\partial \sigma_z}{\partial\theta} = 0, \qquad \frac{\partial \sigma_r}{\partial\theta} = 0, \tag{67}$$

To summarize, if $u = f(r)$, $w = f(z)$, and $v = 0$, then the stresses and strains are independent of θ and z, provided the cylinder is stress free in its initial unloaded state. A square element in an r,z plane, with sides parallel to the r and z axes, becomes a rectangle with sides having the same orientation. The rotations are zero, as can be seen from the following formulas[3].

$$\begin{aligned}
\omega_r &= \frac{1}{2}\left(\frac{1}{r}\frac{\partial w}{\partial\theta} - \frac{\partial v}{\partial z}\right) \\
\omega_\theta &= \frac{1}{2}\left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial r}\right) \\
\omega_z &= \frac{1}{2}\left(\frac{\partial v}{\partial r} + \frac{v}{r} - \frac{1}{r}\frac{\partial u}{\partial\theta}\right).
\end{aligned} \tag{68}$$

---

3. *Love, A.E.H., A Treatise on the Mathematical Theory of Elasticity, Dover Publications, New York, 1944. Page 56.*

On referring to Eqs. (1), (2) and (3) we see that

$$\omega_r = 0, \qquad \omega_\theta = 0, \qquad \omega_z = 0 \tag{69}$$

The type of displacement which occurs in the r,z plane, in both the elastic and plastic zones, is shown in Figure 2.

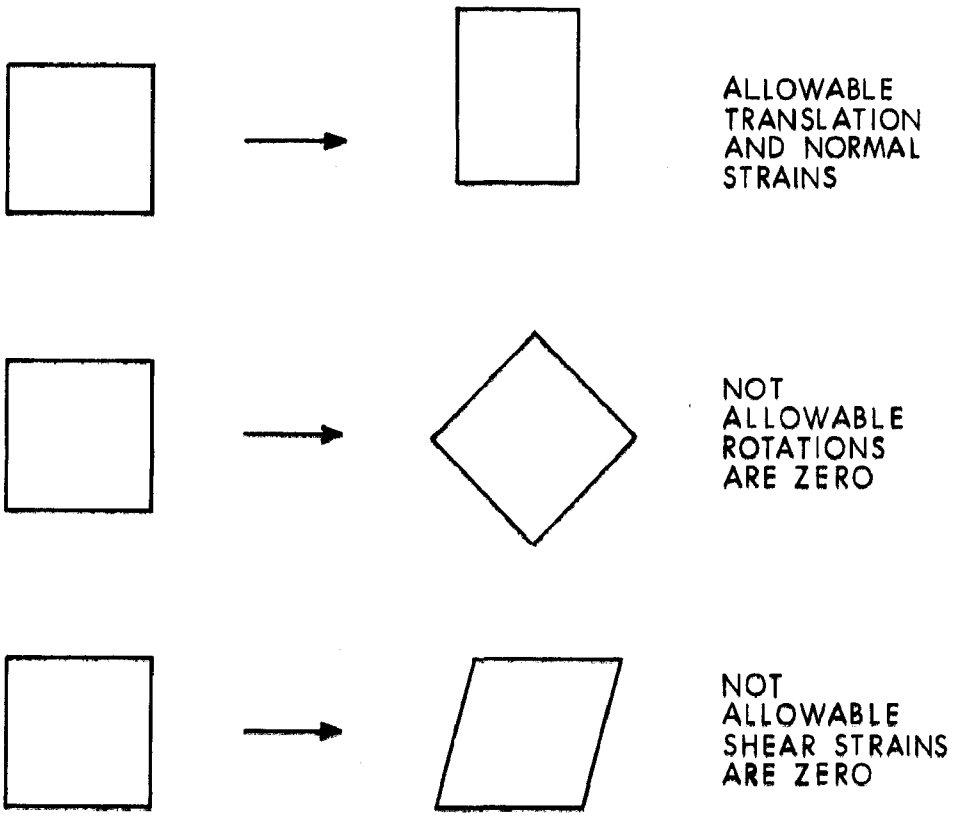


FIGURE 2. Deformation in a Pressurized Elastic Perfectly Plastic Cylinder

We have analyzed the general nature of the stresses and strains in some detail, as this aspect of the autofrettage process was the subject of some discussion during an investigation of a premature in a 175mm gun.

IV. DERIVATION OF THE BASIC EQUATIONS

We now show that the Prager-Hodge equations for plastic flow in a cylinder under plane strain conditions may be modified to account for a non-zero axial strain. Solution of the revised flow equations does not lead to a complete solution of the problem; in essence, we must determine conditions at the elastic-plastic interface that will give

the correct axial load. An iterative procedure is required, since the axial load cannot be determined until the integration of the flow equations has been completed.

The following considerations enter into the derivation of the basic equations:

1) Compatibility relations, leading to the Prager-Hodge strain relation.

2) Equilibrium conditions.

3) Compressibility relation.

4) Von Mises yield condition.

5) Prandtl-Reuss flow equations.

6) Continuity relations at the elastic-plastic interface, $r=\rho$.

7) End conditions involving the axial strain or axial load.

Items 6 and 7 are closely related; the need to satisfy both conditions when the axial load is prescribed leads to an iterative procedure, as the axial load cannot be calculated until integration of the flow equations has been completed.

The compatibility relation for the strains is expressed in terms of the Prager-Hodge strain function and the axial strain deviation. We have

$$\phi = \frac{1}{3}(\varepsilon_r - \varepsilon_\theta)$$

$$\phi = \frac{1}{3}\left(\frac{\partial u}{\partial r} - \frac{u}{r}\right)$$

$$\frac{\partial \phi}{\partial r} = \frac{1}{3}\left(\frac{\partial^2 u}{\partial r^2} - \frac{1}{r}\frac{\partial u}{\partial r} + \frac{u}{r^2}\right)$$

$$e = \frac{1}{3}(\varepsilon_r + \varepsilon_\theta + \varepsilon_z)$$

$$e_r = \varepsilon_r - e, \quad e_\theta = \varepsilon_\theta - e, \quad e_z = \varepsilon_z - e \tag{70}$$

$$e_z = \frac{1}{3}(2\varepsilon_z - \varepsilon_r - \varepsilon_\theta) \tag{71}$$

$$\frac{\partial e_z}{\partial r} = \frac{1}{3}\left(-\frac{\partial^2 u}{\partial r^2} - \frac{1}{r}\frac{\partial u}{\partial r} + \frac{u}{r^2}\right)$$

and finally

$$\frac{\partial \phi}{\partial r} + \frac{\partial e_z}{\partial r} = -\frac{2\phi}{r} \tag{72}$$

The strain deviations e_θ and e_r can be expressed in terms of e_z and ϕ. We find

$$e_\theta = -\frac{1}{2}(3\phi + e_z) \tag{73}$$

$$e_r = \frac{1}{2}(3\phi - e_z) \tag{74}$$

These relations are required in the derivation of the flow equation.

The equilibrium equation is written in terms of stress and strain deviations. We have

$$s = \frac{1}{3}(\sigma_\theta + \sigma_r + \sigma_z) \tag{75}$$

$$\sigma_r = s + s_r, \qquad \sigma_\theta = s + s_\theta, \qquad \sigma_z = s + s_z \tag{76}$$

so that Eq. (21) becomes

$$\frac{\partial s_r}{\partial r} + \frac{\partial s}{\partial r} + \frac{s_r - s_\theta}{r} = 0 \tag{77}$$

But

$$s = 3Ke \tag{78}$$

$$\frac{\partial s}{\partial r} = 3K\frac{\partial e}{\partial r}$$

or

$$\frac{\partial s}{\partial r} = -3K\frac{\partial e_z}{\partial r} \tag{79}$$

We eliminate s_θ by using the yield condition. We note that

$$s_z = -s_\theta - s_r \tag{80}$$

so that Eq. (58) becomes

$$s_\theta^2 + s_\theta s_r + s_r^2 = k^2$$

$$s_\theta = -\frac{1}{2} s_r \pm s_K$$

where for brevity we have written

$$s_K = \sqrt{4k^2 - 3s_r^2}$$

The ambiguous sign is determined by conditions at the inner surface during incipient plastic yielding. Since the stresses are continuous across the elastic-plastic interface the Lame' formulas are correct. Let

a = inner radius of cylinder

ρ = radius of the elastic plastic interface

b = outside radius

$\sigma_r = -p_0$, $\sigma_z = \sigma_0$ at $r=a$

$\sigma_r = -p_1$, $\sigma_z = \sigma_1$ at $r=\rho$

$\sigma_r = 0$, $\sigma_z = \sigma_2$ at $r=b$

We note that

$$\sigma_1 = \sigma_2$$

since the tube is elastic in the zone $\rho<r<b$. The Lame' formulas give

$$\sigma_\theta = \frac{p_1(b^2+\rho^2)}{b^2-\rho^2} \quad \text{at } r=\rho \tag{81}$$

$$\sigma_\theta = \frac{p_0(b^2+a^2)}{b^2-a^2} \quad \text{at } r=a \tag{82}$$

We consider the state of stress for incipient plastic yielding at $r=a$. On substituting the appropriate values of σ_r, σ_θ, and σ_z into Eq. (58), we find

$$\sigma_0^2 (b^2-a^2) - 2\sigma_0 p_0 (b^2-a^2) + p_0^2 (3b^4+a^4) = 3k^2 (b^2-a^2)^2 \tag{83}$$

$$\sigma_0 = \left\{ a^2 p_0 \pm \sqrt{3[k^2 (b^2-a^2) - b^4 p_0^2]} \right\} / (b^2-a^2) \qquad (84)$$

Eq. (83) represents an ellipse in Cartesian coordinates, so there are two permissible values of σ_0 for each value of p_0 unless the quantity under the radical sign is zero. Prager and Hodge have shown the minus sign should be used in Eq. (84) under plane strain conditions. Under generalized plane strain, open end conditions

$$\sigma_0 = 0 \text{ when } r = a,$$

the minus sign again prevails. Under closed end conditions, the pressure of the hydraulic fluid against the end caps produces an axial tension stress in the cylinder.

$$\sigma_0 = a^2 p_0 / (b^2-a^2) \text{ when } r = a$$

so the quantity under the radical sign is zero. We assume the minus sign is correct under all conditions of interest when

$$a < \rho \leq b.$$

We can now determine the conditions at the elastic-plastic interface from an assumed interface pressure p_i.

$$s_z = -2\sqrt{3[k^2 (b^2-\rho^2)^2 - p_1^2 b^4 / 3(b^2-a^2)}} \qquad (85)$$

$$s_r = -\frac{1}{2} s_z - b^2 p_i / (b^2-\rho^2) \qquad (86)$$

$$s_\theta = -\frac{1}{2} s_z + b^2 p_i / (b^2-\rho^2) \qquad (87)$$

We note that the axial stress deviation is zero or negative at the elastic-plastic interface. We assume these conditions are also true in the plastic zone. Then

$$s_z = -\frac{1}{2} s_r - \frac{1}{2} s_k \qquad (88)$$

$$s_\theta = -\frac{1}{2} s_r + \frac{1}{2} s_k \qquad (89)$$

$$s_\theta + s_r + s_z = 0,$$

as required. On substituting the preceding value of s_θ into Eq. (77), we obtain

$$\frac{\partial s_r}{\partial r} - 3K \frac{\partial e_z}{\partial r} = \frac{-3s_r+s_k}{2r} \tag{90}$$

for the equation of equilibrium.

In writing the Prandtl-Reuss flow equations for the stress and strain deviators in the plastic zone, we use ρ rather than t as the independent variable. This substitution is permissible since ρ is a monotonic, increasing function of t during the first loading cycle. We assume, in summary, that plastic yielding and flow are independent of the rate of loading; this is true to a first approximation provided the rate of loading is sufficiently small. First, we eliminate λ from the flow equations

$$2G \frac{\partial e_r}{\partial \rho} = \frac{\partial s_r}{\partial \rho} + \lambda s_r$$

$$2G \frac{\partial e_\theta}{\partial \rho} = \frac{\partial s_\theta}{\partial \rho} + \lambda s_\theta,$$

obtaining the equation

$$2G \left(\frac{\partial e_r}{\partial \rho} s_\theta - \frac{\partial e_\theta}{\partial \rho} s_r \right) = \frac{\partial s_r}{\partial \rho} s_\theta - \frac{\partial s_\theta}{\partial \rho} s_r$$

Now

$$e_r = \frac{1}{2} (3\phi - e_z)$$

$$e_\theta = \frac{1}{2} (-3\phi - e_z)$$

and

$$s_\theta = -\frac{1}{2} s_r + \frac{1}{2} s_k$$

On eliminating e, e_θ, and s_θ from the flow equations, we obtain

$$G(3s_r - s_k) \frac{\partial e_z}{\partial \rho} + (3s_r + 3s_k) \frac{\partial \phi}{\partial \rho} = \frac{4k^2}{s_k} \frac{\partial s_r}{\partial \rho} \tag{91}$$

where, as before, we let

$$s_k = \sqrt{4k^2 - 3s_r^2}$$

The preceding analysis differs from the original work of Prager and Hodge mainly in the use of e_z rather than e as one of the three dependent variables, and in the indeterminate conditions at the elastic-plastic interface. These interface conditions are finally determined by iteration so that one of the following load conditions at the end of the cylinder is satisified:

$\varepsilon_z = 0$ Plane Strain (Prager and Hodge)

$\int_a^b r\sigma_z dr = 0$ Generalized plane strain, open end condition.

$\int_a^b r\sigma_z dr = a^2 p_i/(b^2-a^2)$ Generalized plane strain, closed end condition.

Equations (72), (90), and (91) form a set of partial differential equations in the independent variables r and ρ and dependent variables ϕ, e_z, and s_r. These equations, the initial conditions, and the boundary conditions at the ends of the cylinder complete the formulation of this problem. The solution is obtained by the method of characteristics, as described in the next section.

V. THE METHOD OF CHARACTERISTICS

The characteristics are found by determining the conditions for which a set of six linear equations in the partial derivatives of e_z, ϕ and s_r are inconsistent. The first three equations have been derived in the preceding section; the second set of three equations is found from differential relations among dr, $d\phi$, and ds_r.

$$\frac{\partial e_z}{\partial r} + \frac{\partial \phi}{\partial r} = -\frac{2\phi}{r}$$

$$3K\frac{\partial e_z}{\partial r} - \frac{\partial s_r}{\partial r} = \frac{3s_r - s_k}{2r} \qquad (92)$$

$$f_1\frac{\partial e_z}{\partial \rho} + f_2\frac{\partial \phi}{\partial \rho} - \frac{\partial s_r}{\partial \rho} = 0 \qquad (93)$$

$$\frac{\partial e_z}{\partial r}dr + \frac{\partial e_z}{\partial \rho}d\rho = de_z \qquad (94)$$

$$\frac{\partial \phi}{\phi r}dr + \frac{\partial \phi}{\partial \rho}d\rho = d\phi \qquad (95)$$

$$\frac{\partial s_r}{\partial r}dr + \frac{\partial s_r}{\partial \rho}d\rho = ds_r \qquad (96)$$

where for brevity we have written

$$f_1 = G[3s_r s_k - s_k^2]/4k^2$$

$$f_2 = G[3s_r s_k + 3s_k^2]/4k^2$$

$$s_k = \sqrt{4k^2 - 3s_r^2}$$

The determinant of the coefficients of the partial derivative is

$$\Delta = \begin{vmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 3K & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & f_1 & f_2 & -1 \\ dr & 0 & 0 & d\rho & 0 & 0 \\ 0 & dr & 0 & 0 & d\rho & 0 \\ 0 & 0 & dr & 0 & 0 & d\rho \end{vmatrix}$$

or

$$\Delta = [f_2 - f_1 - 3K]\, dr(d\rho)^2$$

and finally

$$\Delta = [(4G-3K)\, k^2 - 3Gs_r^2]\, dr(d\rho)^2/k^2 \qquad (97)$$

The factor (4G-3K) is generally negative for steel, so the entire bracketed expression is negative. Hence, if

$$\Delta = 0,$$

then dr=0 or dρ=0. The net of orthogonal lines $r=c_r$, $\rho=c_\rho$, where the constants c_r and c_ρ are arbitrary, form the characteristics for this problem. Certain partial derivatives may be discontinuous across these characteristics.

Along the characteristics, the partial differential equations are replaced by ordinary differential equations.

$$\frac{de_z}{dr} + \frac{d\phi}{dr} = -\frac{2\phi}{r} \text{ along } \rho = c_\rho \tag{98}$$

$$3k\frac{de_z}{dr} - \frac{ds_r}{dr} = \frac{3s_r - s_k}{2r} \text{ along } \rho = c_\rho \tag{99}$$

$$f_1\frac{de_z}{d\rho} + f_2\frac{d\phi}{d\rho} - \frac{ds_r}{d\rho} = 0 \text{ along } r = c_r \tag{100}$$

These equations were solved by a first order predictor-corrector method, starting at the elastic-plastic interface and integrating toward the inner radius of the cylinder. The predictor equations are

$$\Delta e_z + \Delta\phi = -\frac{2\phi(r)}{r}\Delta r \tag{101}$$

$$3K\Delta e_z - \Delta s_r = \frac{3s_r(r) - s_k(r)}{2r}\Delta r \tag{102}$$

$$f_1(\rho)\Delta e_z + f_2(\rho)\Delta\phi - \Delta s_r = \Delta\rho \tag{103}$$

and the corrector equations are

$$\Delta e_z + \Delta\phi = -\frac{1}{2}\left[\frac{2\phi(r)}{r} + \frac{2\phi(r+\Delta r)}{r+\Delta r}\right]\Delta r \tag{104}$$

$$3K\Delta e_z - \Delta s_r = \frac{1}{2}\left[\frac{3s_r(r) - s_k(r)}{2r} + \frac{3s_r(r+\Delta r) - s_k(r+\Delta r)}{2(r+\Delta r)}\right]\Delta r \tag{105}$$

$$\frac{1}{2}\left[f_1(\rho) + f_1(\rho+\Delta\rho)\right]\Delta e_z + \frac{1}{2}\left[f_2(\rho) + f_2(\rho+\Delta\rho)\right]\Delta\phi - \Delta s_r = \Delta\rho. \tag{106}$$

The predictor equations were used once and the corrector equations twice for each set of nodal points involved. The range of integration for each independent variable was divided into at least twenty equal intervals.

Integration of the differential equation for the net axial load was started at the inner cylindrical surface using initial conditions appropriate for incipient plasticity.

Closed End Condition: $p_0 = k(b^2 - a^2)/b^2$ (107)

Open End Condition: $p_0 = \sqrt{3}\, k(b^2 - a^2)/\sqrt{a^2 + 3b^2}$ (108)

Plane Strain Condition: $$p_0 = \sqrt{3}\, k(b^2-a^2)/\sqrt{(1-2\nu)a^2+3b^2} \qquad (109)$$

The generalized plane strain condition applies to the closed end and open end examples.

Calculations for the plane strain condition were carried out in the manner indicated by Prager and Hodge, and are included for purposes of comparison. Once the radius ρ of the elastic-plastic interface is chosen, the remaining conditions at the elastic-plastic interface are determined, and integration toward the inner radius can be carried out without difficulty.

Under generalized plane strain conditions, the pressure at the elastic-plastic interface is determined by an iterative procedure. First, a trial value of the interface pressure is selected, generally by extrapolation from previous results, then the axial stress and other quantities are calculated in sequence as shown below.

$$r = \rho_1 \text{ the elastic-plastic interface} \qquad (110)$$

$$\sigma_r = -p_1, \text{ an estimated trial value} \qquad (111)$$

$$\sigma_z = \sigma_1$$

$$h(b,\rho,p_1) = \sqrt{3[k^2(b^2-\rho^2)^2 - p_1^2 b^4]}$$

$$\sigma_1 = (p_1^2\rho^4+3b^4p_1^2-3k^2)/(b^2-\rho^2)(p_1\rho^2+h) \qquad (112)$$

$$\sigma_\theta = p_1(b^2+\rho^2)/(b^2-\rho^2) \qquad (113)$$

$$s_r = (-3b^2p_1+h)/3(b^2-\rho^2) \qquad (114)$$

$$\phi = -2(1+\nu)b^2p_1/3(b^2-\rho^2)E \qquad (115)$$

$$e_z = -2(1+\nu)h/3E(b^2-\rho^2) \qquad (116)$$

The preceding values of s_r, ϕ, and e_z were used as starting values in integration of Eqs. (101), (102), and (103).

The normal stresses and strains for each nodal point were calculated concurrently. First, we calculate the axial strain at the elastic-plastic interface; according to our assumptions, the axial strain is constant across the entire cross section.

$$\varepsilon_z = [(1-2\nu)\rho^2 p_1 - h]/E(b^2-\rho^2), \tag{117}$$

$$a \leqslant r \leqslant b$$

$$\varepsilon_\theta = \varepsilon_z - \frac{3}{2} e_z - \frac{3}{2} \phi \tag{118}$$

$$\varepsilon_r = \varepsilon_z - \frac{3}{2} e_z + \frac{3}{2} \phi \tag{119}$$

$$e = \varepsilon_z - e_z \tag{120}$$

$$s = 3Ke \tag{121}$$

$$s_k = \sqrt{4k^2 - 3s_r^2} \tag{122}$$

$$s_z = -\frac{1}{2} s_r - \frac{1}{2} s_k$$

$$s_\theta = -\frac{1}{2} s_r + \frac{1}{2} s_k$$

$$\sigma_r = s+s_r, \qquad \sigma_\theta = s+s_\theta, \qquad \sigma_z = s+s_z$$

Finally, the axial load was obtained by integration. In the elastic region,

$$\sigma_z = \sigma_1, \qquad \rho \leqslant r \leqslant b$$

and the plastic region the axial stress was given by the sequence of equations given above. The axial load is given by

$$F = 2\pi \int_a^b r\sigma_z \, dr$$

$F = 0$ in the open end condition

$$F = \pi p_i a^2 \text{ in the closed end condition}$$

If F was not correct, a new interface pressure was chosen and the entire sequence of calculations repeated. The reguli falsi method for obtaining the correct value of p_i worked well for the open end condition[4,5].

Under all end conditions we have

$$p_i \leq k(b^2-\rho^2)/b^2;$$

otherwise we have a negative quantity under the radical sign in the definition of k. The equality sign holds under closed end conditions when ρ equals a; consequently, we find successive approximations for p_i obtained by standard methods did not always satisfy the above inequality. Moreover, the results were greatly affected by round-off error. To overcome these difficulties, the successive approximations for p_i were forced to satisfy the above inequality, and the results were calculated to a high degree of precision. Details of the method of calculation finally adopted will be discussed in a forthcoming report.

VI. RESULTS AND CONCLUSIONS

Extensive numerical and graphical results were obtained for a range of wall ratios, for both plane strain and generalized plane strain. Results for a wall ratio two, open and closed end conditions, under generalized plane conditions are presented in this paper. Additional results, including the effects of varying Poisson's ratio, will be presented in a forthcoming report.

The results for the open end condition are presented in Figures 3 through 9. The existence of axial stresses is clearly indicated. We notice that the axial stress has a maximum within the plastic zone after the elastic-plastic interface has progressed about halfway through the tube. This result was unexpected, yet appears to be con-

---

4. *Hill, R., The Mathematical Theory of Plasticity, Oxford, 1956. Pages 112, 113.*

5. *Hoffman, O. and Sachs, G., Introduction to the Theory of Plasticity for Engineers, McGraw Hill Book Company, 1953. Page 92.*

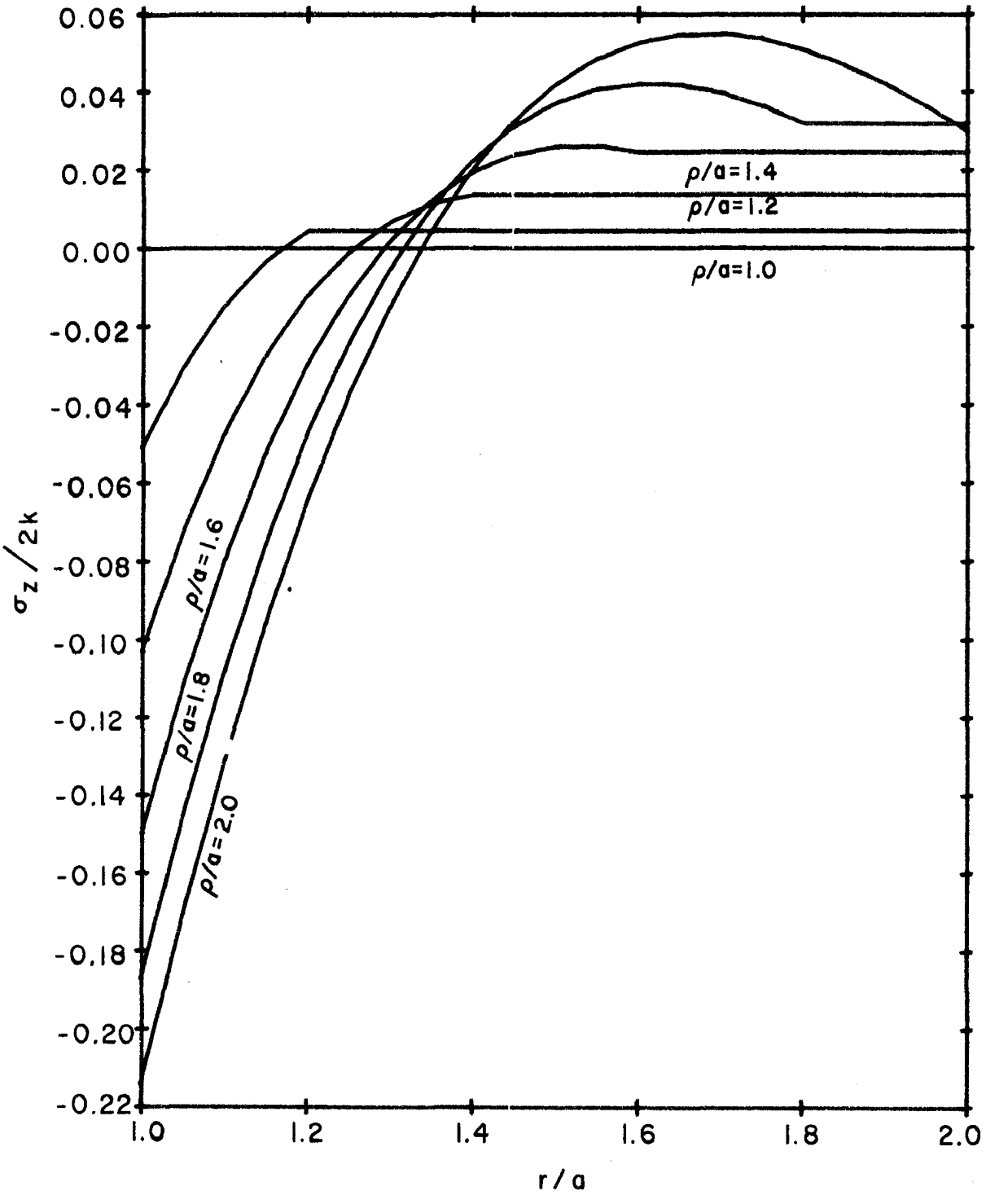


FIGURE 3. Axial Stress Distribution Open End, Generalized Plane Strain Condition

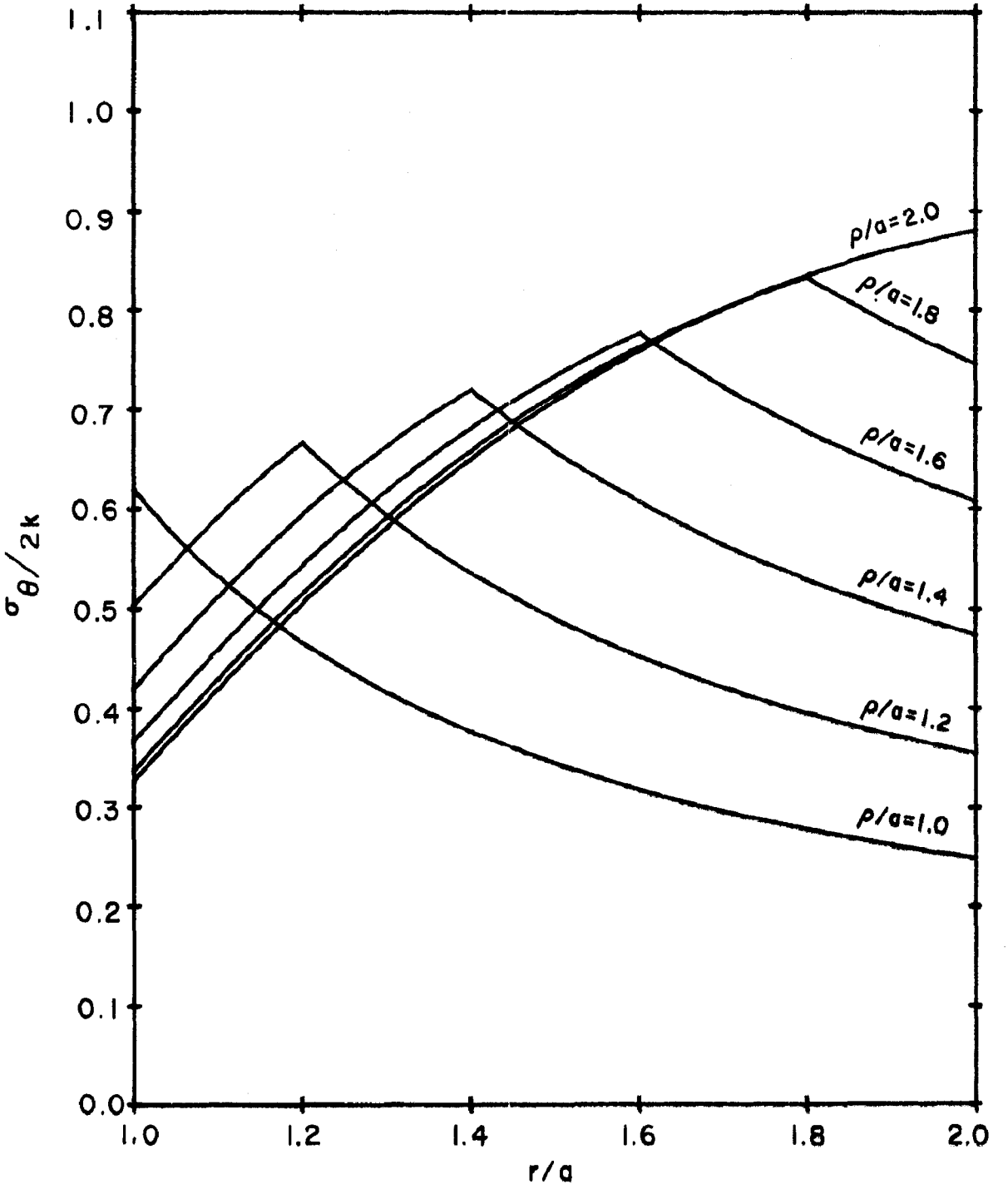


FIGURE 4. Circumferential Stress Distribution, Open End, Generalized Plane Strain Condition

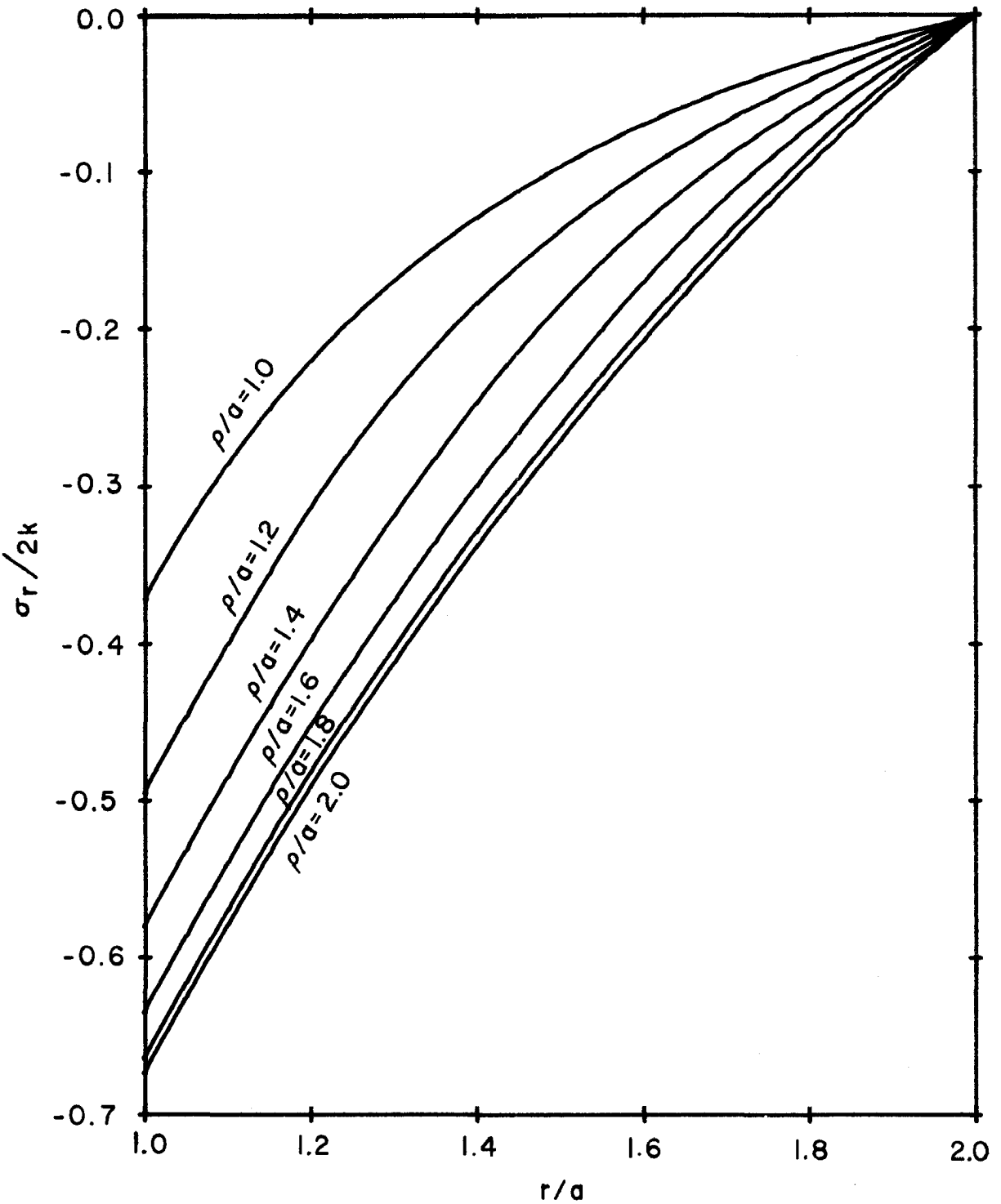


FIGURE 5. Radial Stress Distribution, Open End, Generalized Plane Strain Condition

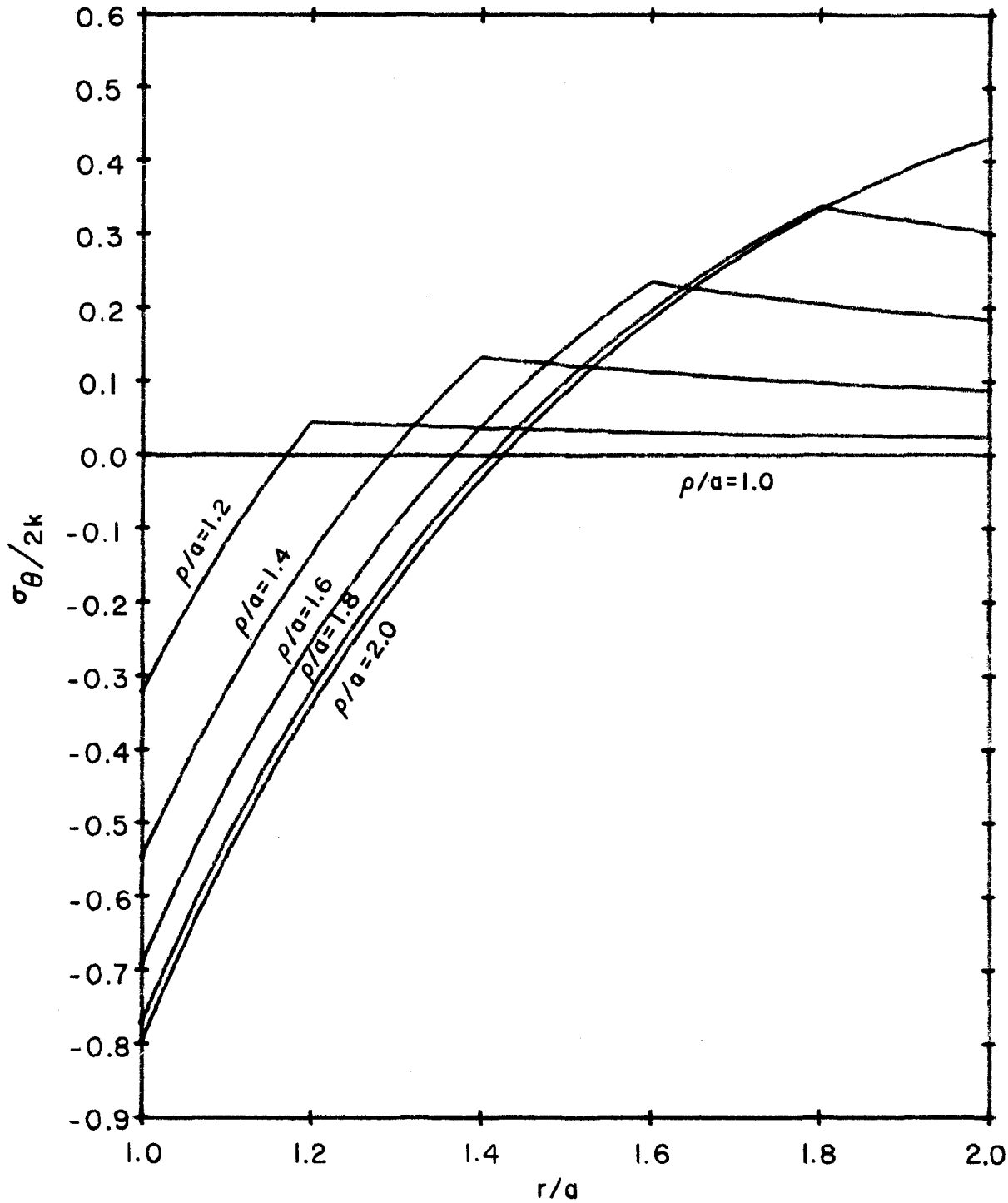


FIGURE 6. Residual Circumferential Stress Distribution, Open End, Generalized Plane Strain Condition

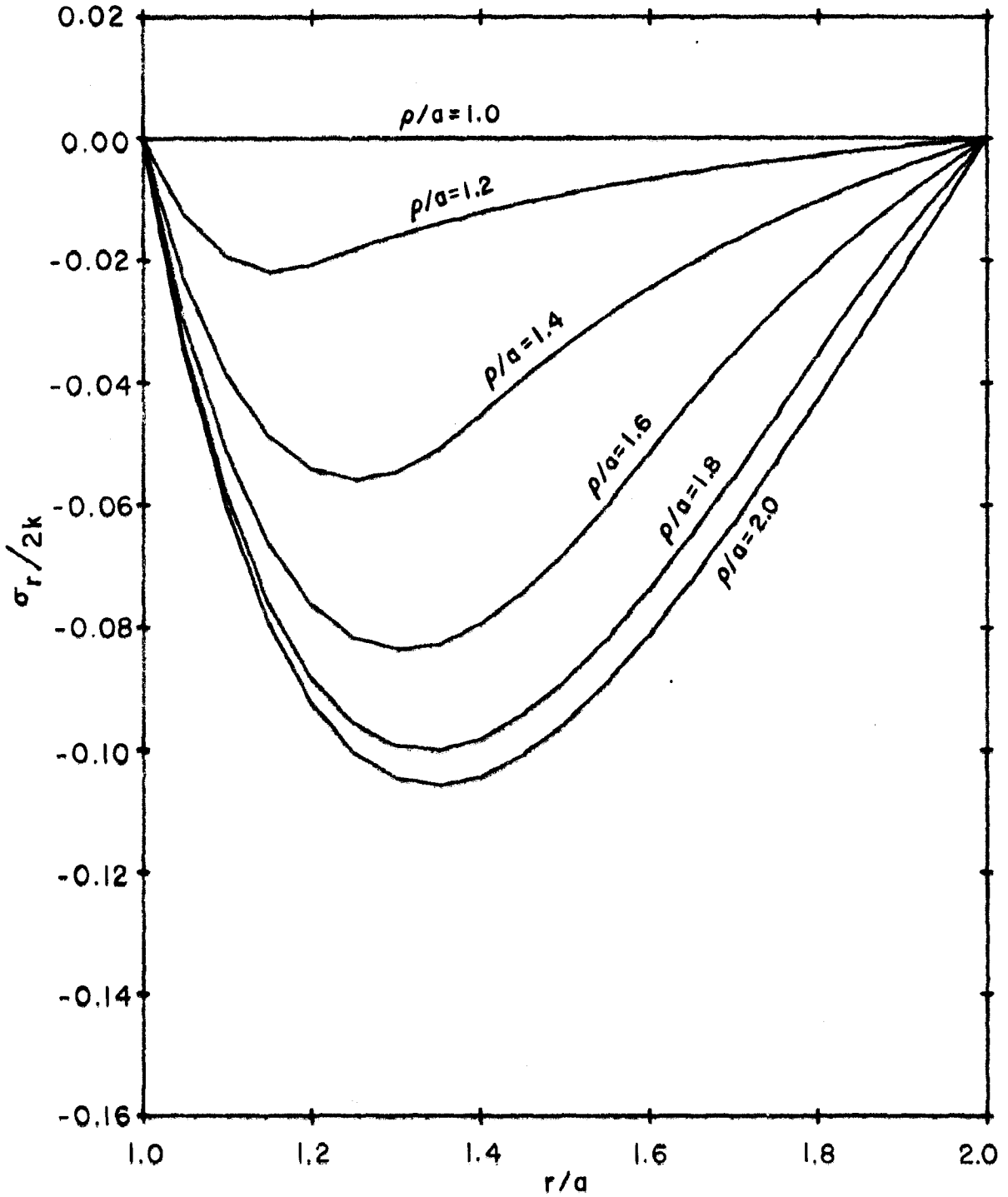


FIGURE 7. Residual Radial Stress Distribution, Open End, Generalized Plane Strain Condition

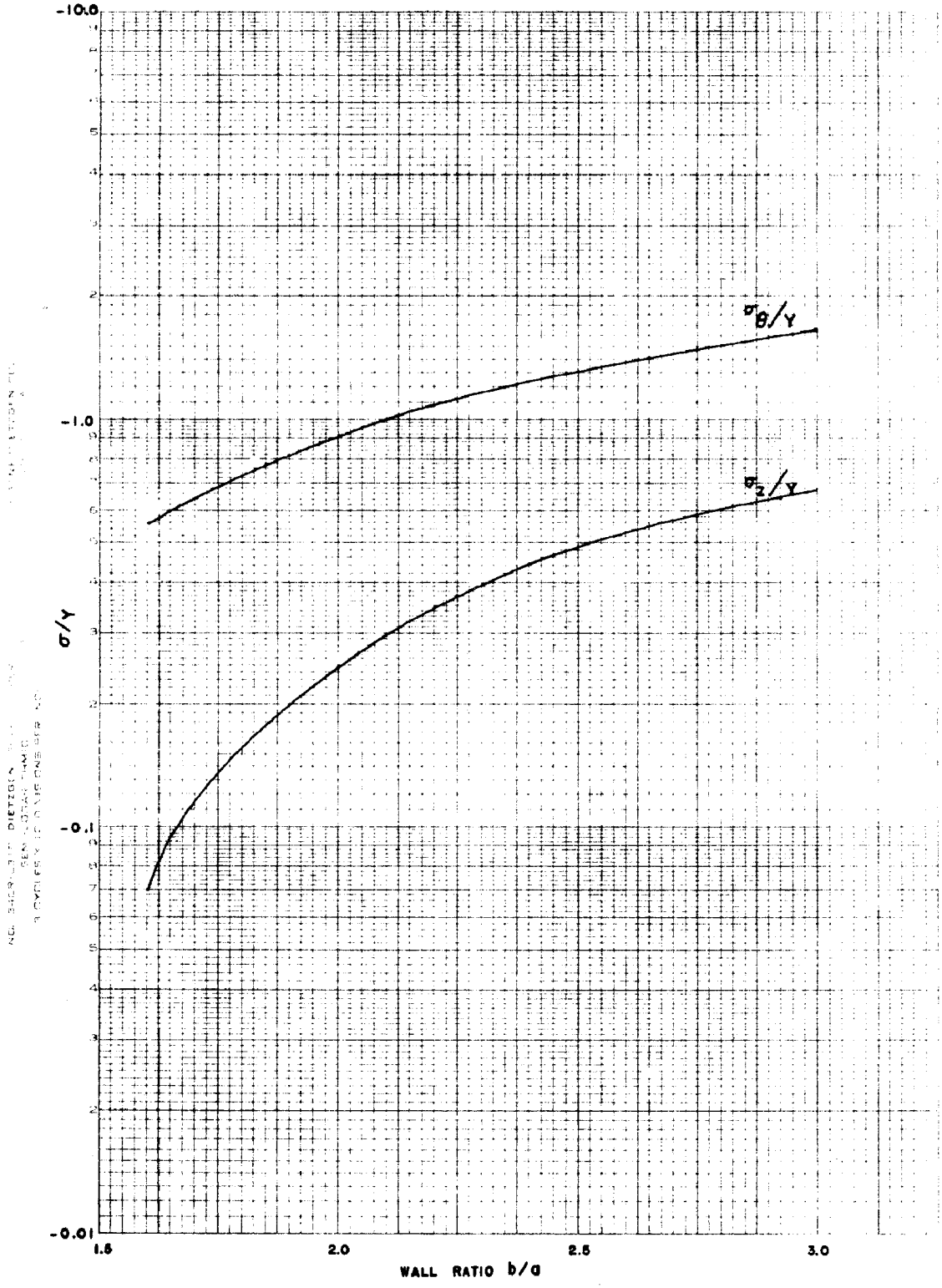


FIGURE 8. Residual Stresses at the Inner Surface as a Function of Wall Ratio, Open End, Generalized Plane Strain Condition

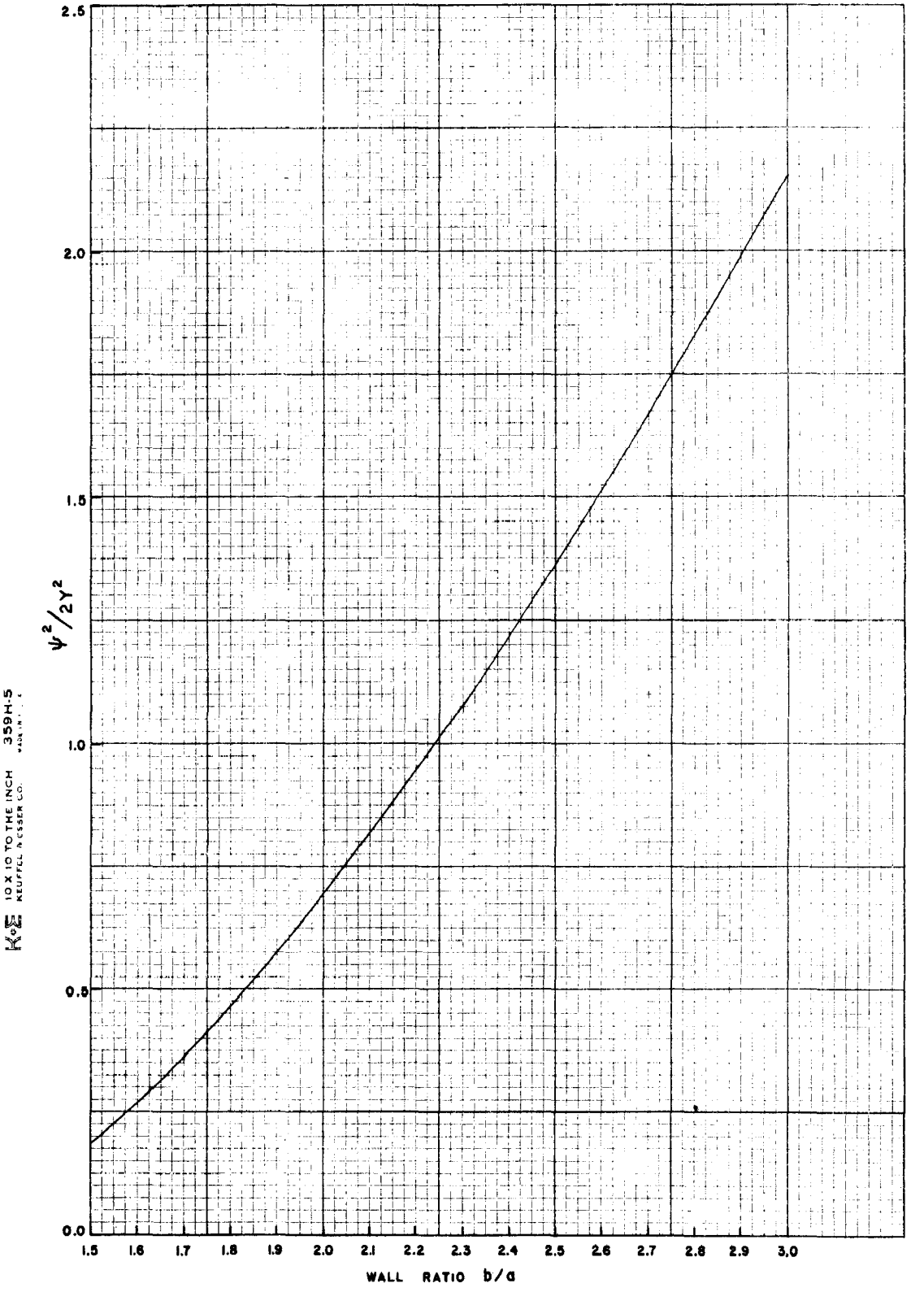


FIGURE 9. Effective Combined Residual Stress as a Function of Wall Ratio, Open End, Generalized Plane Strain Condition

firmed by recent calculations based on the finite element method[6]. The residual axial stress is tensile and equals about one fourth of the residual circumferential stress for the wall ratio under consideration.

Similar calculations for open end conditions were carried out for wall ratios in the range 1.5-3.0 inclusive. The corresponding residual stresses were also calculated, assuming the re-yielding would not occur at the inner surface; these results are shown in Figure 8. The effective compressive stresses corresponding to the Von Mises yield condition were also calculated to determine the maximum wall ratio for which the effective compressive residual stress at the inner surface was less than the yield stress; after several iterations a wall ratio of 2.25 was obtained, as shown in Figure 9. If the entire cylinder yields plastically when pressurized, then re-yielding will occur at the inner surface for larger wall ratios; the dotted lines indicate predicted residual stresses in the absence of re-yielding. A new analysis and a corresponding program would be required to calculate the actual stress conditions under re-yield conditions.

Calculations for the closed end condition were also carried out, as shown in Figures 10 through 14. The pressure on the end caps of the cylinder produces a large tensile stress in the cylinder. For this reason the axial stress is larger than that obtained for the open end condition, and increases steadily through the thickness of the plastic zone. In actual practice, we would expect the end conditions would be intermediate between the open and closed end conditions, as friction in the seals would contribute to the axial stress. The analysis of this paper could be modified to take friction into account if we assumed a constant coefficient of friction for all values of the internal pressure.

Experiments to confirm the predictions of the preceding analysis and calculations would be desirable. It is essential that the length to diameter ratio be large enough to prevent end effects from influencing the stress and strain distribution in the central part of the cylinder, where presumably these quantities would be measured. Measurements in both the circumferential and axial directions are required in order to fully determine the stress distribution on the outside of the cylinders. Special instrumentation is required to obtain measurements of the strain at the inner surface due to the presence of oil under high pressure. The Sachs method of measuring residual stresses can be modified to measure axial as well as circumferential residual stresses.

---

6. *Chen, P.C.T., The Finite Element Analysis of Elastic-Plastic Thick-Walled Tubes, Watervliet Arsenal Technical Report WVT-74039, September, 1974. Page 23, Figure 5. This report contains an excellent bibliography of recent work, with emphasis on current research at Watervliet Arsenal.*

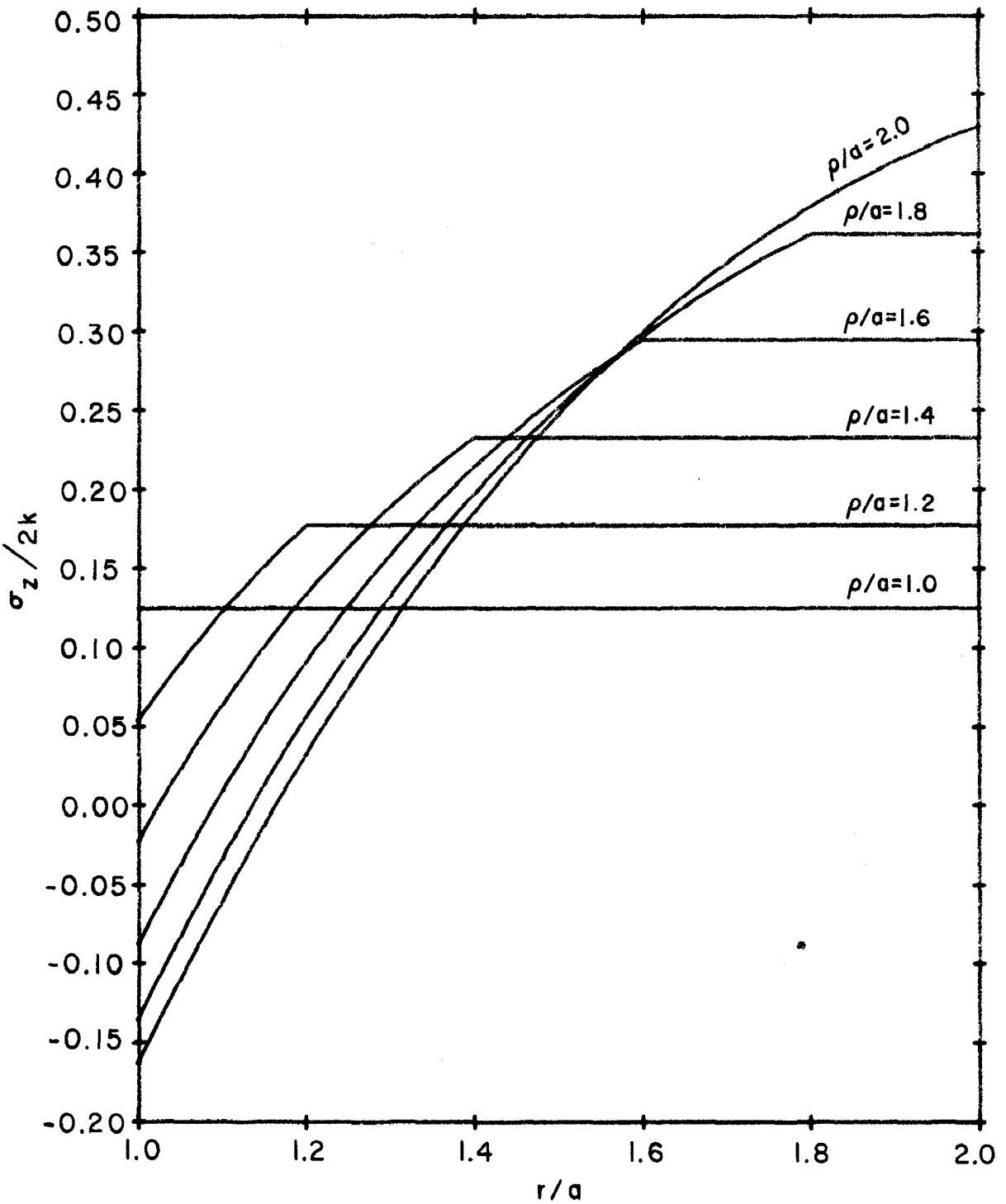


FIGURE 10. Axial Stress Distribution, Closed End, Generalized Plane Strain Condition

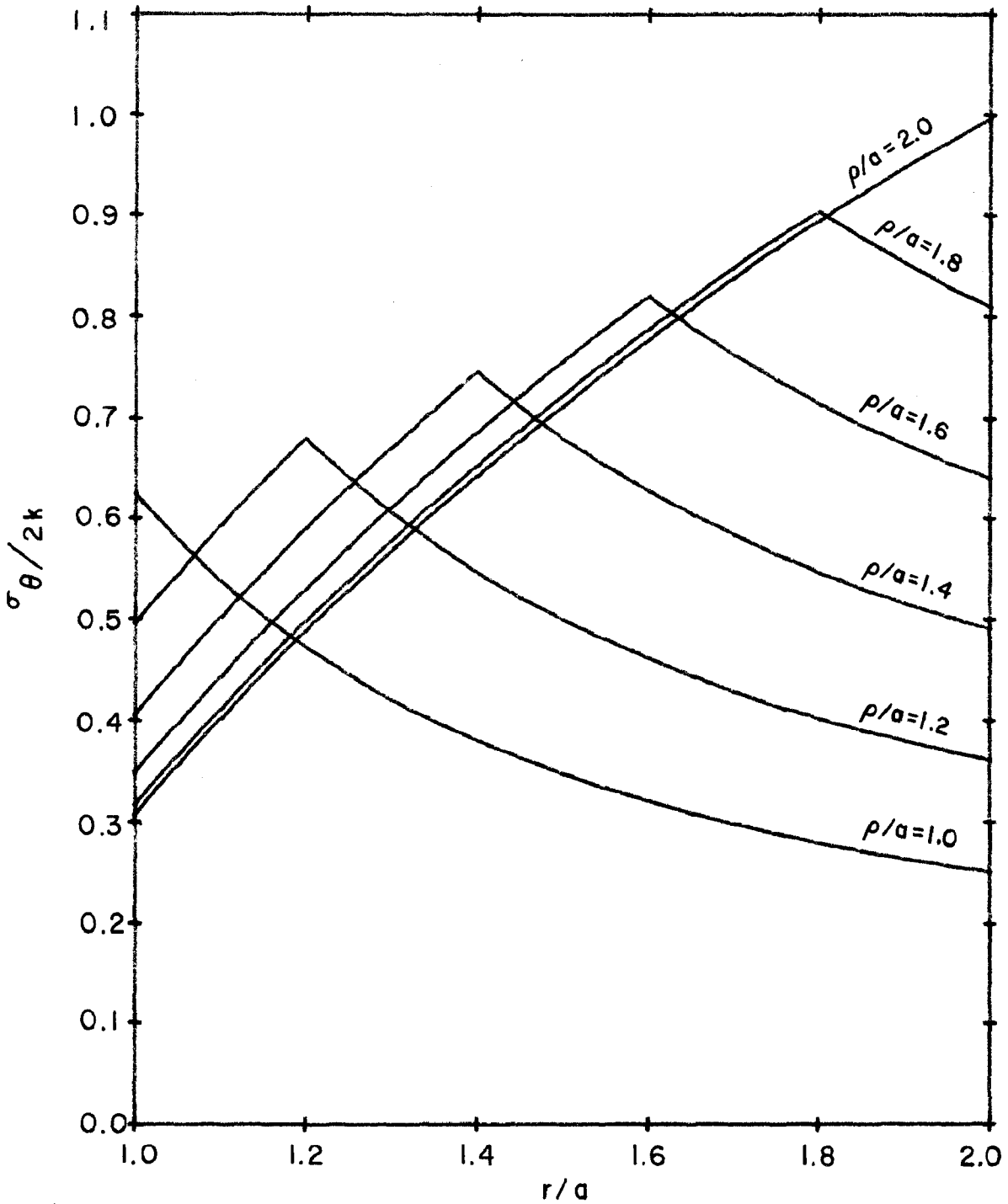


FIGURE 11. Circumferential Stress Distribution, Closed End, Generalized Plane Strain Condition

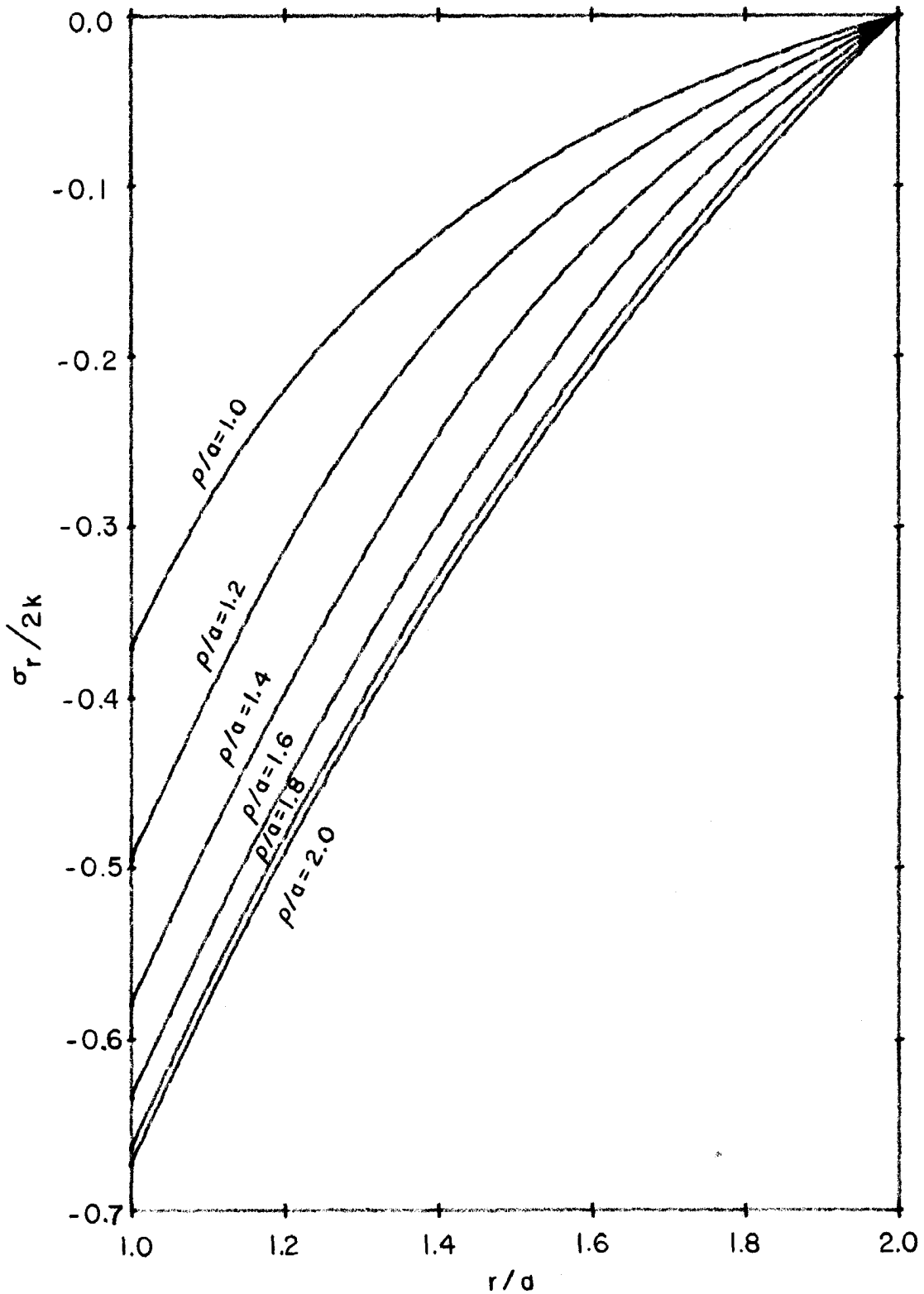


FIGURE 12. Radial Stress Distribution, Closed End, Generalized Plane Strain Condition

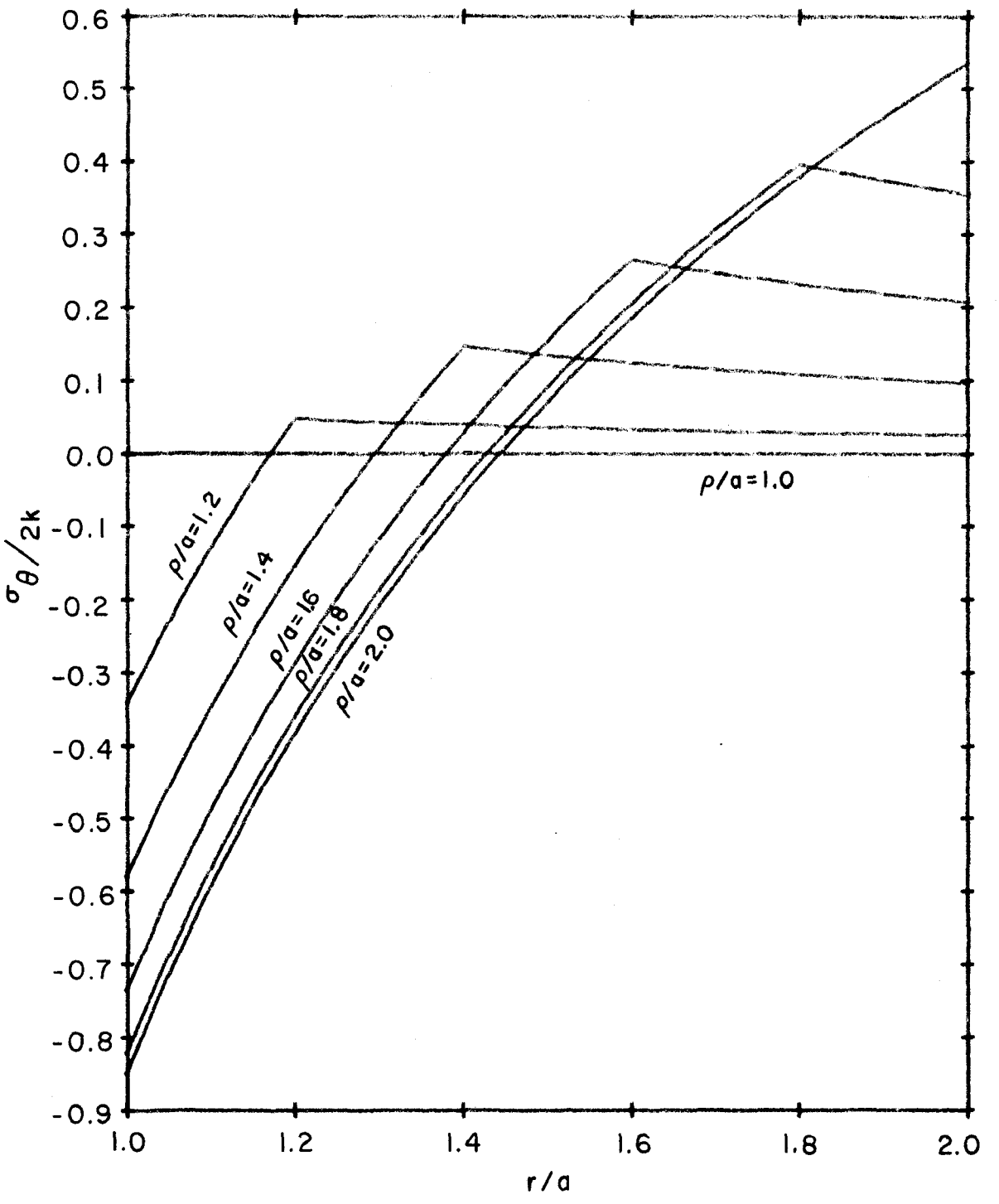


FIGURE 13. Residual Circumferential Stress Distribution, Closed End, Generalized Plane Strain Condition

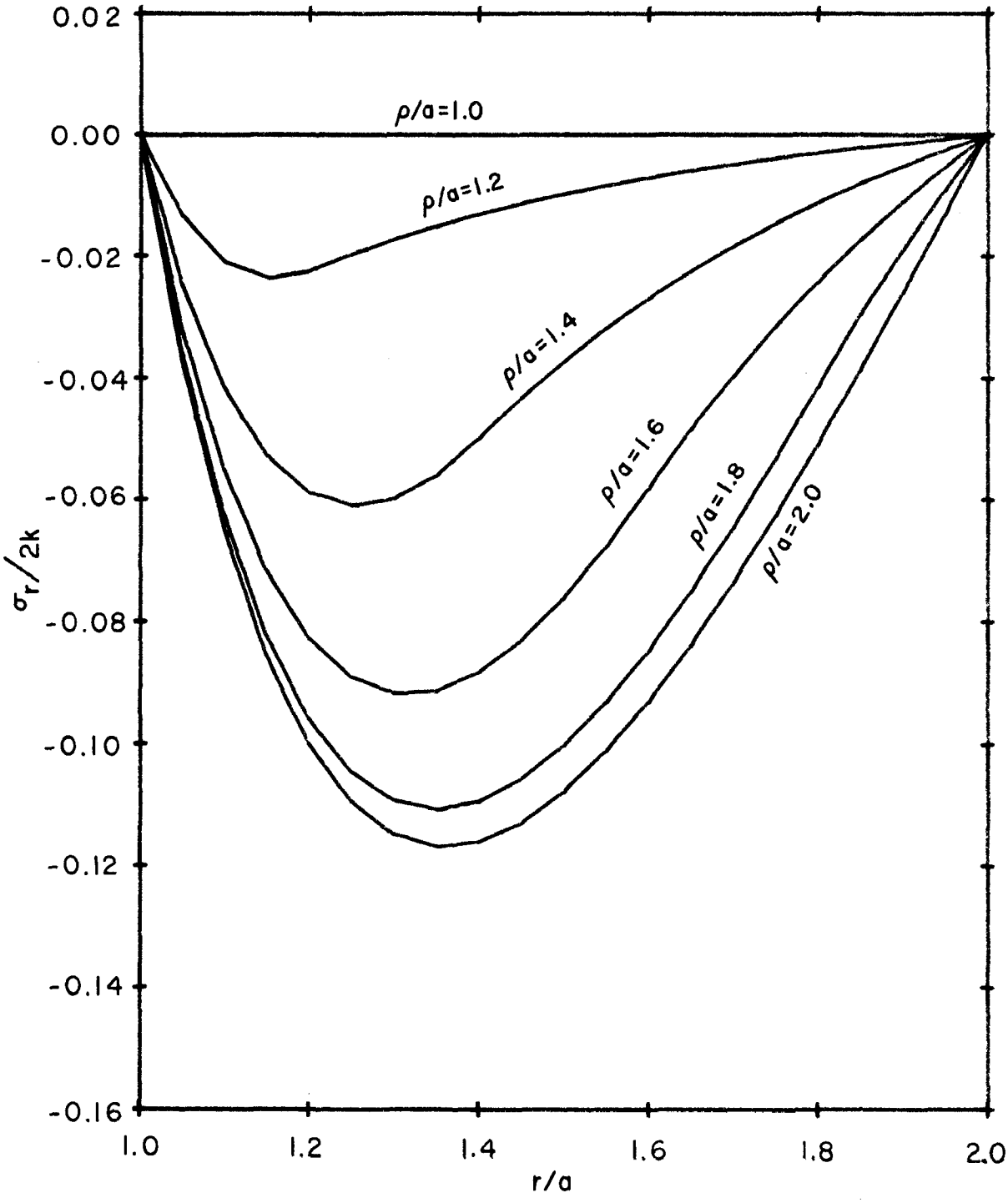


FIGURE 14. Residual Radial Stress Distribution, Closed End, Generalized Plane Strain Condition

In calculating the residual stresses, it has been assumed the material is elastic and the stress-strain relations are linear during the release of pressure. However, some non-linearity may be expected as the re-yield condition is approached, due to the Bauschinger effect. For this reason the actual residual effective stress may be somewhat less than the calculated value. It is hoped the analysis and calculations in this paper will make it possible to delineate the respective roles of the end conditions and the Bauschinger effect in producing the observed field of residual stresses.

ACKNOWLEDGMENTS

I wish to acknowledge the assistance of Dr. B.P. Burns in verifying the basic equations, of Dr. J. Frasier for helpful comments on the theory of characteristics, of Dr. J. Giese for a discussion of linear equations and the symmetry of the stresses, and of LT J. Carson for locating the source of the numerical difficulty with the closed end condition.

REFERENCES

1. Timoshenko, S. and Goodier, I.N., Theory of Elasticity, second edition, McGraw Hill Book Company, Inc., New York, 1951. Pages 305, 306.

2. Prager, W. and Hodge, Jr., P.G., Theory of Perfectly Plastic Solids, Dover Publications, New York, 1968. Pages 16-32, 95-122.

3. Love, A.E.H., A Treatise on the Mathematical Theory of Elasticity, Dover Publications, New York, 1944. Page 56.

4. Hill, R., The Mathematical Theory of Plasticity, Oxford, 1956. Pages 112, 113.

5. Hoffman, O. and Sachs, G., Introduction to the Theory of Plasticity for Engineers, McGraw Hill Book Company, 1953. Page 92.

6. Chen, P.C.T., The Finite Element Analysis of Elastic-Plastic Thick-Walled Tubes, Watervliet Arsenal Technical Report WVT-74039, September, 1974. Page 23, Figure 5. This report contains an excellent bibliography of recent work, with emphasis on current research at Watervliet Arsenal.

NONLINEAR PROBLEMS IN CHEMICALLY REACTING DIFFUSIVE SYSTEMS

Donald S. Cohen
Department of Applied Mathematics
California Institute of Technology
Pasadena, California 91125

ABSTRACT

Various problems occuring in chemical reactor theory and the theory of chemically or biochemically reacting mixtures are studied. In particular, we investigate the processes controlling multiplicity and its implications with regard to ignition and extinction in a reactor and the processes by which stable oscillatory states are set up. In inhomogeneous systems with spatially and temporally distributed parameters many of the phenomenia are locally distributed. Various perturbation procedures have been developed to analytically study these problems and to simplify numerical procedures.

1. Introduction. We shall formulate and describe certain recently occurring problems arising in various fields such as the theory of chemical and biochemical reactions, chemical and nuclear reactors, combustion, diffusion through membranes and porous media, Joule heating, and soil mechanics. We do not wish to imply that any one field, much less all of them, is treated globally. In fact, we wish to strongly emphasize the point that although the equations describing the various problems have a certain common form, the parameters and specific nonlinearities are different, and the specifics of each problem clearly determine the techniques used and the results obtained. It is just not enough to think formally about a class of problems of which the one of interest is a special case. Thus, although we present the general class of equations for the above-mentioned fields, we shall present results only for two specific problems in an attempt to give some idea of the type of techniques used and the kinds of results obtained. A rather complete history of recent work as well as references to the work of many researchers can be found in the references [1]-[5].

Mathematically, in the simplest one-dimensional geometries all the problems consist of finding the temperature $T(x, t)$ and the concentrations $C_i(x, t)$, $i = 1, \ldots, N$, of the reacting species at any point x at any time t in the region of interest. Allowing for chemical reaction, molecular diffusion, and convection, the governing equations for either adiabatic or non-adiabatic situations are given by

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(D_0 \frac{\partial T}{\partial x}\right) - \alpha_0 \frac{\partial T}{\partial x} + F_0 (C_1, \ldots, C_N\ T) \quad , \tag{1}$$

$$\frac{\partial C_i}{\partial t} = \frac{\partial}{\partial x}\left(D_i \frac{\partial C_i}{\partial x}\right) - \alpha_i \frac{\partial C_i}{\partial x} + F_i (C_1, \ldots, C_N, T) \quad , \quad i = 1, \ldots, N-1 \ . \tag{2}$$

Here the D_k represent the diffusion coefficients (which in some problems could depend explicitly on the C_k and/or T), α_k are known constants, and the F_k are almost always of the form of polynomials in the concentrations C_k multiplied by the usual exponential Arrhenius temperature dependence (or sums of such forms). We shall confine our attention to the one-dimensional situation described by (1), (2). The necessary modification for a multi-dimensional problem is clear; namely, the diffusive terms become $\mathrm{div}(D_k \,\mathrm{grad}\cdot\,)$ and the convection terms involve the appropriate gradients. Note that we need only the equations for (N - 1) concentrations since the usual stochiometric equation (mass balance) easily gives the concentration of the remaining species once all the others are known.

In actual reaction problems three situations <u>all</u> occur which make simplification of the system of equations (1), (2) possible. (i) Some of the equations completely decouple from the rest (by themselves or in blocks). (ii) Three time scales usually predominate so that some compounds are formed much faster or much slower than the scale of interest. Thus, some compounds are formed or destroyed so fast that they can always be considered in equilibrium, and some compounds can be taken to be constant (often at non-equilibrium values) because they are so slowly reacting and thus simply undergo a slow drift in the value of the concentration. (iii) Various dimensionless parameters occur in widely disparate sizes. Thus, a very successful approach has been to combine singular perturbation methods for both large and small values of various parameters with multi-time scale perturbation methods. For problems in the theory of both continuous stirred tank and tubular chemical reactors A. B. Poore [6] has found with various numerical checks that the

perturbation analysis for large Peclet number $(P \to \infty)$ provides good results for P as low as 2 or 3, and the analysis for small Peclet number $(P \to 0)$ provides good results for P close to unity. Thus, it is believed that the perturbation analysis yields all possible phenomena which can occur.

To illustrate the diverse phenomena which can occur we will describe one set of operating characteristics for a simple non-adiabatic tubular reactor for one range of the parameters.

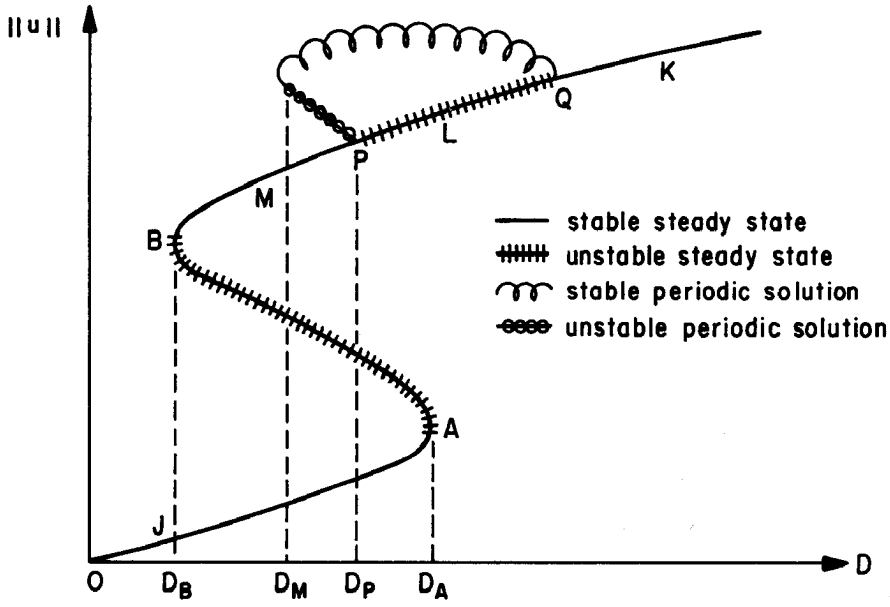


Fig. 1

Figure 1 is the response diagram (i.e., the behavior of the amplitude of the temperature $\|T\| = \max_{0 \le x \le 1} [T]$ as a certain parameter, the Dahmkohler number D, varies) for values of the physical parameters in certain ranges (and these ranges are given explicitly by A. B. Poore [6]). We shall trace the process as D is increased starting at a small amplitude stable steady state for D near 0. The response moves along the path 0JA.

As D is increased past D_A, the temperature undergoes a jump to a large amplitude stable periodic response above point L, and as D is further increased, the amplitude of this oscillation decreases until the oscillation vanishes at point Q. The stable steady state is followed up the branch through K as D increases still further. Now, as D is decreased, the process follows the branch KQLMB, the solution being steady and stable from K to Q and from M to B with a stable oscillation of increasing amplitude from Q to M. As D decreases through D_M, the oscillation just ceases (subcritical branching at P). As D is decreased below D_B, there is an extinction as the response jumps to point J and then follows the path J to 0. The analysis and numerical computations to support this description are given in [7]-[8].

For the continuous stirred tank reactor and the simple tubular reactor it has been possible to identify fourteen or fifteen different response diagrams depending on the various ranges of the six independent physical parameters. Briefly stated, many of the outstanding problems consist of finding all the possible response diagrams for more involved reactors such as adiabatic and non-adiabatic packed bed and moving bed reactors perhaps involving higher order kinetics. An extremely interesting implication of some of these distributed type reactors is the appearance of solutions involving interior and moving boundary layers [10].

To illustrate the types of phenomena currently arising in problems involving so-called localized temporal and spatial instabilities we shall consider the following equations:

$$\frac{\partial u}{\partial t} = A - (B+1)\,u + u^2 v + D_1 \frac{\partial^2 u}{\partial x^2} \quad , \tag{3}$$

$$\frac{\partial v}{\partial t} = Bu - u^2 v + D_2 \frac{\partial^2 v}{\partial x^2} \quad . \tag{4}$$

Here B is a parameter and A(x) is a given function of x. These equations (together with appropriate boundary and initial conditions) have been proposed by I. Prigogine (see [11], [12] for references) as a model to describe observed localized temporal and spatial structures involving concentration waves and localized instabilities in chemically reacting systems. The equations (3), (4) arise in writing the conservation laws for a chemical reaction $A + B \rightarrow D + E$ through a certain autocatalytic step. u and v represent the chemical concentrations of two reactants, and once they are found the concentrations of all other products and reactants are found by solving certain linear parabolic problems by routine methods.

Perhaps the simplest presentation of some of the phenomena contained in this system can be made by means of Figures 2 to 9 which have been taken from [11]. The (D_2, B)-plane is divided into three regions as shown in Fig. 2.

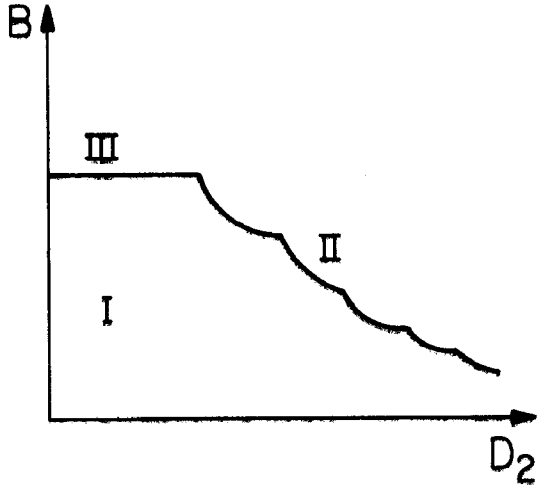


Fig. 2

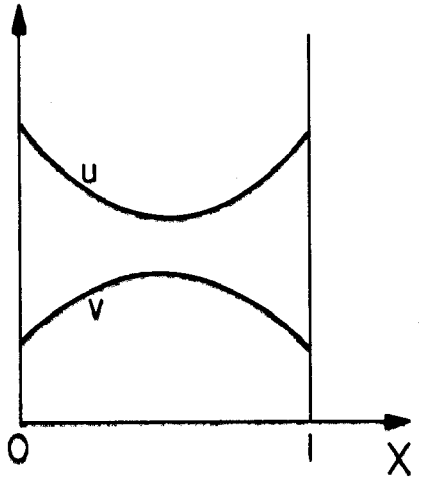


Fig. 3

For fixed D_2, B in region I there exists a unique steady state solution v, v of the problem, and these solutions have the form illustrated in Fig. 3. As the values of D_2 and B are changed so that we cross

from region I into region II, these solutions become unstable and the new stable steady states of Fig. 4 evolve. These new states clearly differ significantly

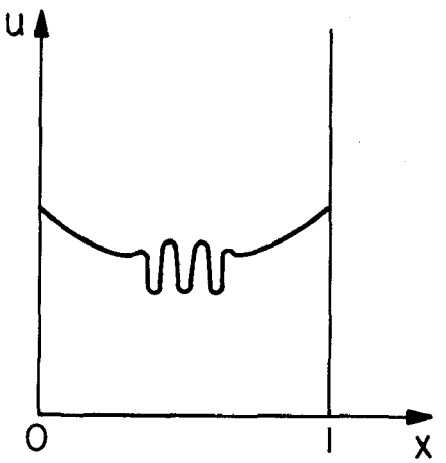


Fig. 4

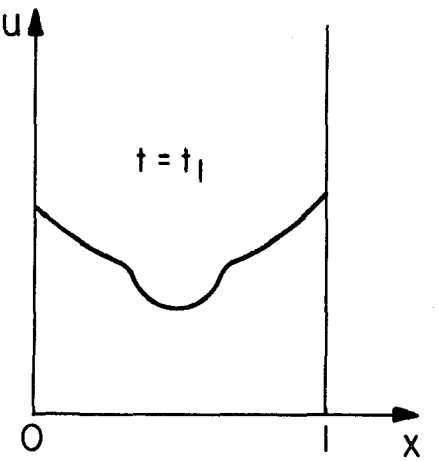


Fig. 5

from the old states only through a localized spatial oscillation. As we cross from region I to region III, the stable steady states of region I lose their stability, and localized stable temporal oscillations (i.e., concentration waves) are set up as illustrated in Figs. 5-8 (which show the behavior of the solution over <u>half</u> a period). For fixed $x = \frac{1}{2}$, Fig. 9 shows that this oscillation is of relaxation type.

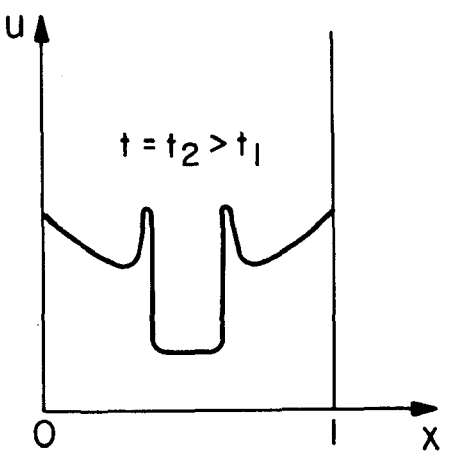


Fig. 6

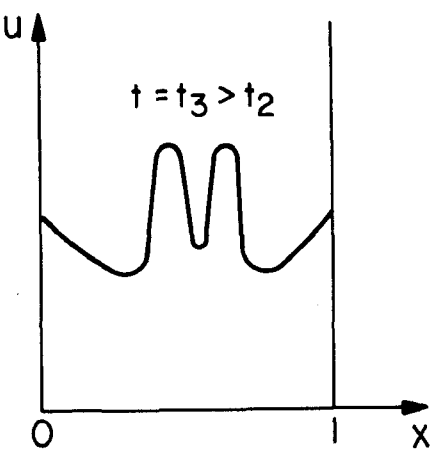


Fig. 7

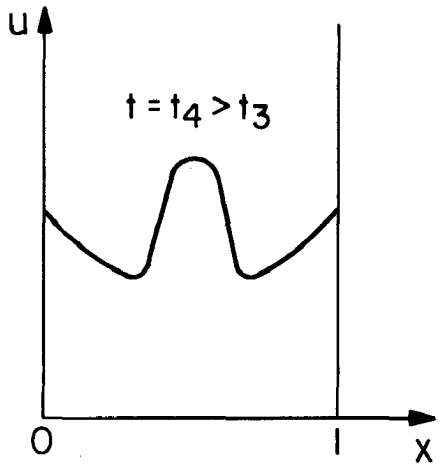


Fig. 8

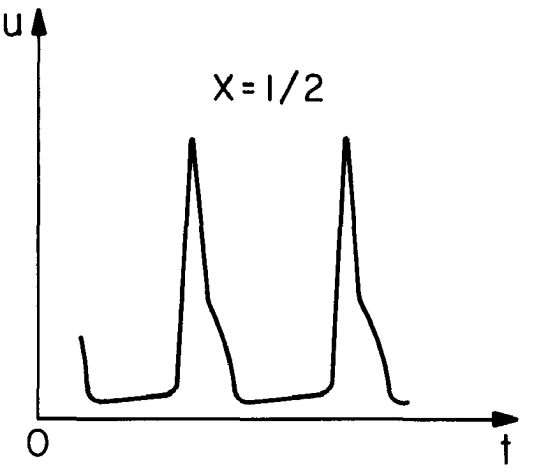


Fig. 9

Thus, as stability boundaries are crossed the new stable states differ from the old stable states only in local regions. This localization is due to the non-uniform distribution A(x) of the reactant A. Mathematically, this phenomenon is suggested by the presence of turning-points in the linearized stability equations. The analytical calculations to support the structure as we have just described it is given by J. A. Boa [12] by using various combinations of WKB, singular perturbation, and multi-time scale perturbation techniques.

The model we have just described is one of the few currently receiving considerable attention in that it is a simple reaction-diffusion system with qualitatively accurate descriptions of certain simple reacting systems. In particular, it is thought that certain aspects of the Belousov-Zhabotinsky reaction (see [13] and [14] for detailed descriptions) can be accounted for with this model. This reaction yields relaxation oscillations and various concentration wave interactions all easily visible to the naked eye as color changes in the reacting chemicals. The results of N. Kopell

and L. N. Howard [14], [15] and J. A. Boa [12] and J. A. Boa and D. S. Cohen [4] account qualitatively for much of the observed phenomena.

Ultimately, it would be desirable to consider the more complex problems in biochemical pattern formation (i.e., the spatial and temporal structure of the evolving systems). Experimental observations are available and some generally accepted mathematical models exist. Of particular current interest are certain questions concerning biochemical oscillations (on scales ranging from a few seconds to several days). A review containing an extensive list of references has been given by B. Hess and A. Boiteux [16]. In many problems the chemical kinetics (for example, the rate functions) and the (possibly nonlinear) diffusion coefficients are not completely known, but their general forms are usually known, and from this it is possible to derive much of the qualitative theory.

References

1. D. S. Cohen, Multiple solutions of nonlinear partial differential equations, lectures in Nonlinear Problems in the Physical Sciences and Biology, Proceedings of a Battelle Summer Institute (Seattle, July, 1972) Springer-Verlag, 1973.

2. D. S. Cohen, Bifurcation from multiple eigenvalues, to appear.

3. D. S. Cohen and J. P. Keener, Multiplicity and stability of oscillatory states in a continuous stirred tank reactor with exothermic consecutive reactions $A \rightarrow B \rightarrow C$, to appear.

4. J. A. Boa and D. S. Cohen, Localized instabilities in reaction diffusion equations, SIAM J. Appl. Math., to appear.

5. R. Aris, The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts; Vol. I: The Theory of The Steady State; Vol. II: Questions of Uniqueness, Stability, and Transient Behavior, Oxford Press, 1974.

6. A. B. Poore, Stability and bifurcation phenomena in chemical reactor theory, Ph.D. thesis, California Institute of Technology, Pasadena, 1972.

7. D. S. Cohen, Multiple solutions and periodic oscillations in nonlinear diffusion processes, SIAM J. Appl. Math., 25 (1973) 640-654.

8. D. S. Cohen and A. B. Poore, Tubular chemical reactors: The "lumping" approximation and bifurcation of oscillatory states, SIAM J. Appl. Math., (1974).

9. C. R. McGowin and D. D. Perlmutter, Tubular reactor steady state and stability characteristics, AICHe J., 17 (1971) 831-849.

10. I. H. Farina and R. Aris, Transients in distributed chemical reactors. Part 2: Influence of diffusion in the simplified model, Chem. Eng. J., 4 (1972) 149-170.

11. M. Herschkowitz-Kaufman and G. Niclois, Localized spatial structures and nonlinear chemical waves in dissapative systems, J. Chem. Phys., 56 (1972) 1890-1895.

12. J. A. Boa, A model biochemical reaction, Ph.D. thesis, California Institute of Technology, Pasadena, 1974.

13. N. Kopell and L. N. Howard, "Plane wave solutions to reaction-diffusion equations," Studies in Appl. Math., 52 (1973).

14. A. M. Zhabotinsky and A. N. Zaikin, "Autowave processes in a distributed chemical system," J. Theor. Biol. 40 (1933) 45-61.

15. L. N. Howard and N. Kopell, "Wave trains, fronts, and transition layers in reaction-diffusion equations," to appear.

16. B. Hess and A. Boiteux, "Oscillatory phenomena in biochemistry," Annul Review of Biochem., 40 (1971) 237-258.

A NEW NUMERICAL METHOD OF SOLUTION OF SCHRODINGER'S EQUATION

George Morales
Army Missle Test and Evaluation Directorate (TE-PC)
White Sands Missle Range, New Mexico 88002

and

Robert G. McIntyre

University of Texas at El Paso
El Paso, Texas 79968
and Instituto de Ciencias Biomedicas
Universidad Autonoma de Ciudad Juarez
Juarez, Chihuahua, Mexico

ABSTRACT. The potential function is approximated in the one dimensional Schrodinger equation by a step function with an arbitrary, finite number of steps. In each step the resulting differential equation has constant coefficients and is integrated exactly in terms of the trigonometric or hyperbolic functions. The solutions are then matched at the interface of each layer. The eigenfunction is then constructed over the entire domain. This numerical method has certain unique features; (a) the potential function does not have to be known analytically; (b) for a given fixed number of steps in the potential approximation, all the eigenfunctions and eigenvalues have the same absolute accuracy; (c) any number of eigenvalues and eigenfunctions can be obtained in a single computer run without any need to guess initial eigenvalues; (d) for a given fixed number of steps in the potential approximation we could obtain the whole infinite spectrum of eigenvalues and eigenfunctions; (e) very low computation time on the computer.

1. INTRODUCTION. A numerical method is presented for the solution of one dimensional Schrodinger equation which is rather good from a practical (computation time) and conceptual point of view.

The potential is approximated by a step function with an arbitrary but finite number of steps. In each step the resulting differential equation has constant coefficients and is integrated exactly in terms of trigonometric or hyperbolic functions. The solutions are then matched at the interface of each layer, and the eigenfunction is then constructed over the entire domain. A very familiar idea in Quantum Mechanics is used in the development of the numerical method which is the matching of the Schrodinger equation solutions and their derivatives at the interface of the square well potentials used and then solving for the constant coefficients. This work represents the implementation and testing of a numerical algorithm which solves the Schrodinger equation for a step potential function with an arbitrary number of steps.

This method has been tested on several problems and the numerical results are very good. The only input into the computer program is a numerical table of the potential. No initial estimates of the eigenvalues are necessary. The computer program in a single pass will output any desired number of eigenvalues and the corresponding eigenfunctions and their nodes. Computation time is roughly equal for the eigenvalues as for the eigenfunctions. The eigenvalues are computed independently from the eigenfunctions and can be computed only by themselves thereby cutting computation time in half if only the eigenvalues are wanted.

In this numerical method the potential is approximated by a step function, but once the approximate Schrodinger equation is set up, it is solved exactly. What this implies is that all eigenvalues and eigenfunctions are all of the same accuracy. The reason for this is that all the eigenfunctions are exact solutions to a given Schrodinger equation (i.e., they are written down explicitly in terms of trigonometric and hyperbolic functions). The numerical results obtained substantiate this expectation.

In more conventional methods such as Rayleigh-Ritz [1], the higher eigenvalues and eigenfunctions are not as accurate as the fundamental eigenvalue and eigenfunction, making it necessary to progressively increase the number of mesh points in order to compute higher eigenvalues and eigenfunctions. The reason for this is that the eigenfunctions are themselves approximations to the solutions of a given Schrodinger equation. In any one of these conventional approximation methods, as the higher eigenfunctions oscillate more rapidly, as their order increases, more mesh points are necessary to compute them with a given accuracy. This does not occur in the method presented in this thesis.

2. STATEMENT OF THE PROBLEM. The one dimensional Schrodinger equation is written in dimensionless form as

$$-\Delta^2 y + V(x)y = Ey$$

or

$$\frac{d^2y}{dx^2} - V(x)y + Ey = 0 \quad , \tag{1}$$

where $V(x)$ is the potential function and E is the energy eigenvalue. For central field problems and for bound states

(a) $V(x)$ is infinite at the origin $x = 0$,
(b) $V(x)$ has a negative minimum value for some $x = a$,
(c) $V(x)$ approaches zero asymptotically as $x \to \infty$.

We could also consider problems for symmetric potential wells with infinitely high walls, because this method is quite general and can be applied in both cases. The Schrodinger equation will be treated as a Sturm-Liouville problem (i.e., the calculations are restricted to a finite domain), and therefore the character of the potential function $V(x)$ is of

secondary importance with respect to the computations. We will restrict Eq. (1) to the following homogeneous boundary conditions

$$y(0) = y(L) = 0 \quad . \tag{2}$$

The boundary conditions (2) are rigorous for a potential well problem with infinitely high walls; for a central field problem it is necessary to approximate the right boundary condition such that the eigenfunctions remain finite for x by taking suitably large L in (2) [2]. This is a good approximation since for large enough L, the eigenfunctions approach infinitely small values.

3. ANALYTICAL ASPECTS. We begin by approximating an analytical potential V(x) such as in Fig. 1, if available, by a step function in a well defined way which need not be specified now, in the following way:

$$V(x) \doteq \begin{cases} V_1 & 0 < x < x_1 \\ V_2 & x_1 < x < x_2 \\ V_3 & x_2 < x < x_3 \\ \vdots & \vdots \\ V_n & x_{n-1} < x < x_n = L \end{cases} \quad . \tag{3}$$

In (3), the constant values $V_1, V_2, \ldots, V_n$ are step function approximations of an analytical potential V(x). The potential can be approximated with arbitrary accuracy $0(h)$, if a sufficiently small step width $h = x_i - x_{i-1}$ is utilized [3]. Therefore, our approximate problem (3) can approach the exact problem as closely as we desire, solely by picking the desired step widths h. The analytical form of the potential need not be known in order to apply the approximation (3). (See Fig. 2.) Once a given number of steps is chosen, the approximate problem is solved exactly in terms of elementary trigonometric and hyperbolic functions.

In each layer i, the approximate problem becomes

$$\frac{d^2y}{dx^2} + (E - V_i)y = 0 \ , \quad i = 1,2,\ldots,n \quad . \tag{4}$$

We shall define

$$a_i \equiv E - V_i \ , \quad b_i^2 \equiv |a_i| \tag{5}$$

so that the solution to (4) is

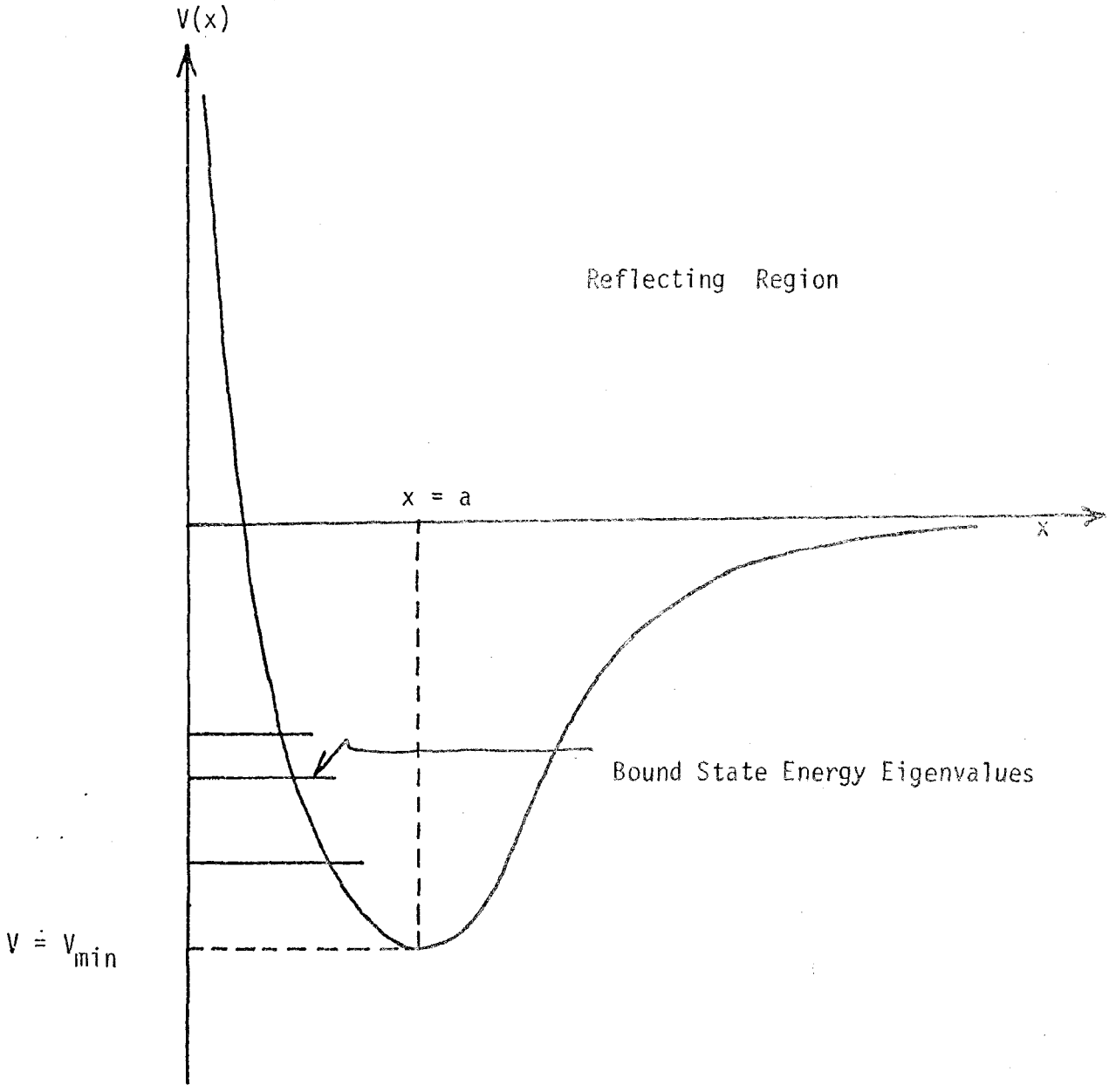


Figure 1. Typical Attractive or Repulsive Potential Curve.

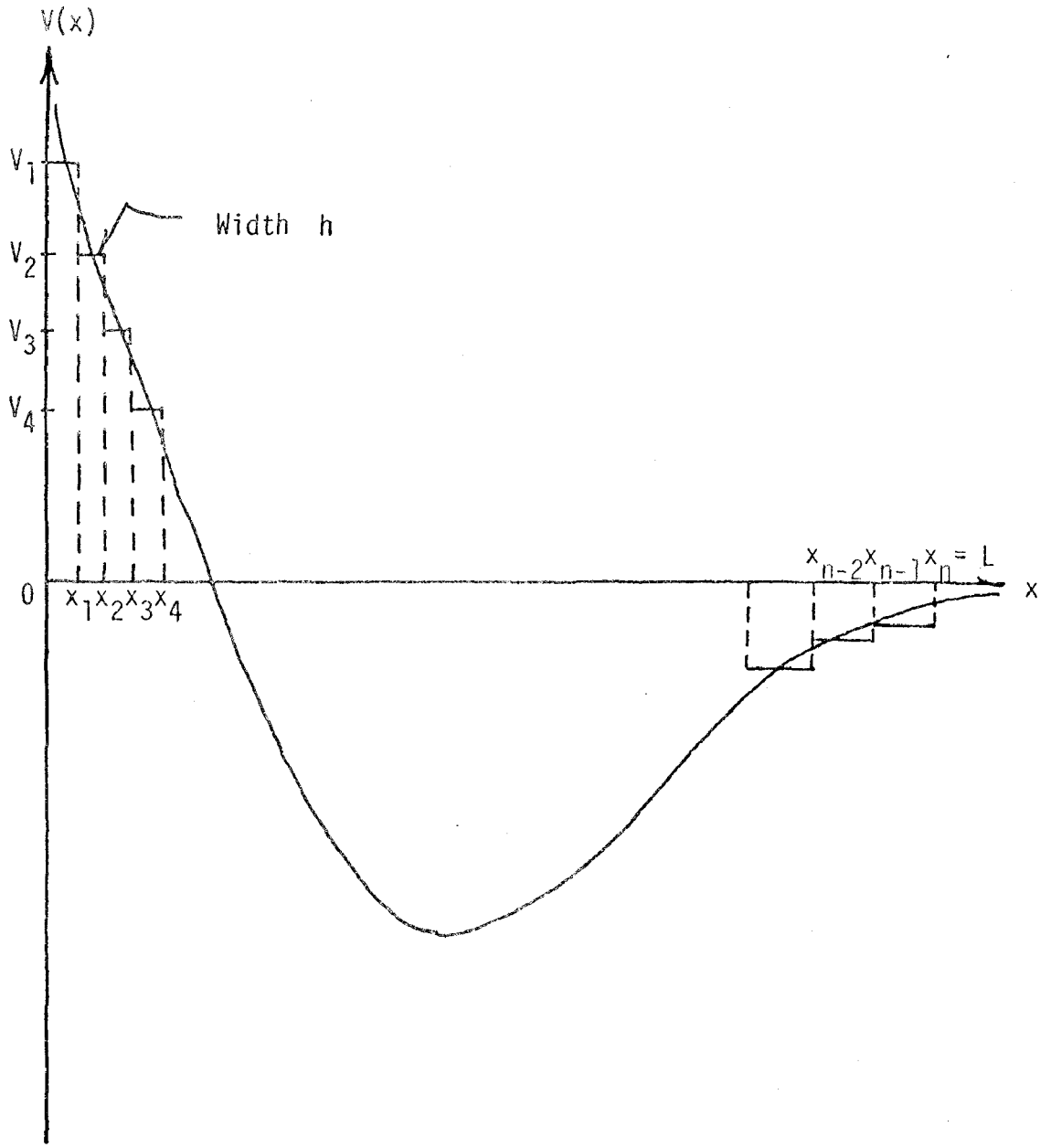


Figure 2. Typical "Stepped" Potential Indicating Step Widths.

(6) $$y = A_i F(b_i x) + B_i G(b_i x) \ , \quad i = 2,3,\ldots,n-1$$

where F and G are given in Table I, and A_i and B_i are integration constants.

Table I

Solutions of Equation (4)

	$a_i < 0$ (Forbidden Region)	$a_i = 0$	$a_i > 0$ (Allowed Region)
$F(b_i x)$	$\cosh(b_i x)$	1	$\cos(b_i x)$
$G(b_i x)$	$b_i^{-1} \sinh(b_i x)$	x	$b_i^{-1} \sin(b_i x)$

These solutions exhibit the well-known fact that the Schrodinger equation solutions are oscillatory within the region defined by the two turning points and exponential outside. In central field problems, the classical turning points are defined by

(7) $$a_i \equiv E - V_i = 0$$

where the total energy equals the potential energy. Classically this is the point at which the incident kinetic energy of the incident particle equals its potential energy, and therefore the point at which the kinetic energy of the particle is zero. It will, at the next instant, change the direction of its motion; therefore these points, at which $E = V_i$, are referred to as "classical turning points." It is interesting to note here that other approximations, namely the W.K.B.J.* approximation, fail in the neighborhood of turning points and require special consideration [4]. Any one of a number of textbooks on Quantum Mechanics will cover the W.K.B.J. approximation [5]. In summary, the W.K.B.J. method is a short wave length approximation technique used for solving the Schrodinger Eq. (1). As the wavelength decreases the variation of the potential V(x) over one wavelength becomes smaller. The approximation is then made that in the limit we may consider the potential V(x) as a constant for several wavelengths about x and that in this region the momentum is

(8) $$p(x) = \sqrt{2m(E - V(x))} \ .$$

The corresponding approximate solution of Eq. (1) is then given by the plane wave solution

(9) $$y = \exp(\pm\, i/h \int p \, dx) \ .$$

*The letters W.K.B.J. stand for G. Wentzel, H. A. Kramers, L. Brillouin, and H. Jeffreys, who more or less independently rediscovered the procedure in connection with the solution of different problems.

The "quasi-classical" condition that the wavelength of the particle vary slowly over distances of the order of itself is written

$$\left|\frac{d(\lambda/2\pi)}{dx}\right| << 1 \quad , \tag{10}$$

where $\lambda(x) = 2\pi h/p(x)$ is the de Broglie wavelength of the particle and

$$\frac{p}{h} = k = \sqrt{\frac{2m(E - V)}{h^2}}$$

is the wave number.

Condition (10) can be rewritten as

$$\left|\frac{d}{dx}\left(h(2m(E - V))^{-1/2}\right)\right| = \left|-\frac{h}{2}(2m(E - V)^{-3/2}\frac{d}{dx}(2m(E - V))\right| \tag{11}$$

$$= \left|\frac{mh}{(2m(E - V))^{3/2}}\frac{dV(x)}{dx}\right| << 1 \quad .$$

It can readily be seen that the solution (9) will fail at those points (the turning points) at which $E = V(x)$, that is, the zeros of Eq. (11). This problem is not encountered in our method of solution.

The solution (6) in regions 1 and n are given respectively by

$$y = A_1F(b_1x) + B_1G(b_1x) \ , \quad y = A_nF(b_nx) + B_nG(b_nx) \quad . \tag{12}$$

Now applying boundary conditions (2) to these solutions, we obtain

$$y = B_1G(b_1x) \ , \quad y = B_nG[b_n(x - L)] \tag{13}$$

since in neither region 1 or n is the function F equal to zero, which in turn implies that

$$A_1 = A_n = 0 \quad .$$

The function G is defined in Table I, and B_1 and B_n are integration constants.

We are now left with the straightforward task of determining the integration constants A_i and B_i. This is done by matching the solutions (6) and its derivatives at the interfaces, namely

(14) $$y_i = y_{i+1} \ , \quad y_i' = y_{i+1}' \ .$$

The general equations for matching the solutions and its derivatives at the ith interface are

(15)
$$A_i F(b_i x_i) + B_i G(b_i x_i) = A_{i+1} F(b_{i+1} x_i) + B_{i+1} G(b_{i+1} x_i) \ ,$$
$$A_i F'(B_i x_i) + B_i G'(b_i x_i) = A_{i+1} F'(b_{i+1} x_i) + B_{i+1} G'(b_{i+1} x_i) \ .$$

At the first interface, which involves boundary region 1 in which $A_1 = 0$, the first matching pair of solution and derivative equations become

(16)
$$B_1 G(b_1 x_1) = A_2 F(b_2 x_1) + B_2 G(b_2 x_1) \ ,$$
$$B_1 G'(b_1 x_1) = A_2 F'(b_2 x_1) + B_2 G'(b_2 x_1) \ .$$

At the second interface the matching pair of solution and derivative equations are

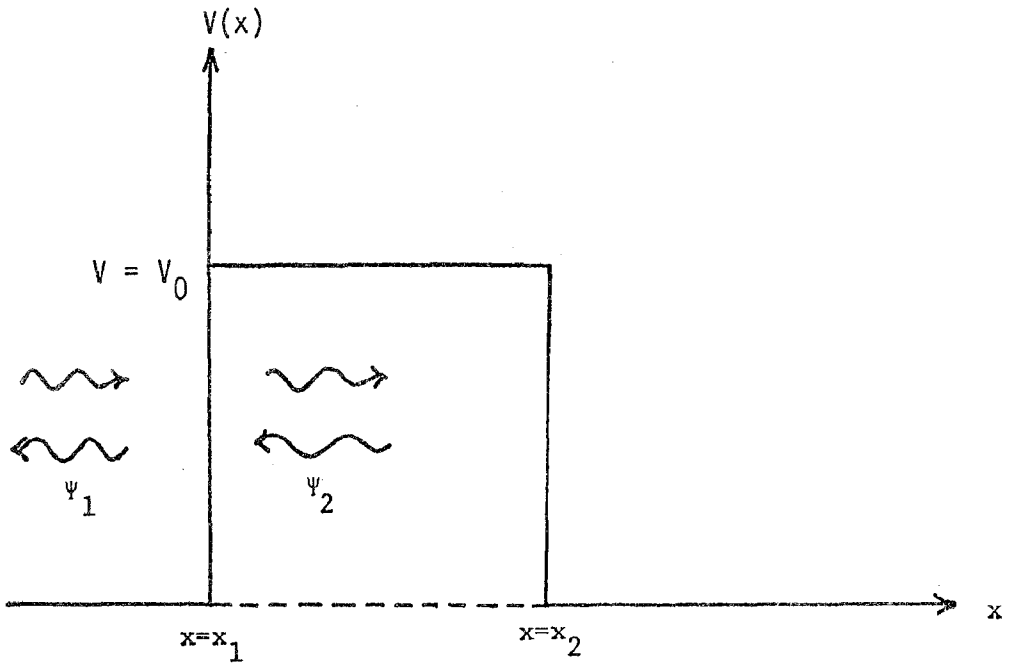
(17)
$$A_2 F(b_2 x_2) + B_2 G(b_2 x_2) = A_3 F(b_3 x_2) + B_3 G(b_3 x_2) \ ,$$
$$A_2 F'(b_2 x_2) + B_2 G'(b_2 x_2) = A_3 F'(b_3 x_2) + B_3 G'(b_3 x_2) \ .$$

This matching process is performed at all interfaces until reaching the interface at x_{n-1} which involves boundary region n in which $A_n = 0$. Therefore, the final pair of equations matching solution and derivative at the rightmost interface x_{n-1} are

(18)
$$A_{n-1} F(b_{n-1} x_{n-1}) + B_{n-1} G(b_{n-1} x_{n-1}) = B_n G[b_n (x_{n-1} - L)]$$
$$A_{n-1} F'(b_{n-1} x_{n-1}) + B_{n-1} G'(b_{n-1} x_{n-1}) = B_n G'[b_n (x_{n-1} - L)] \ .$$

Rewriting this system of equations in a homogeneous form, we obtain

(19)
$$B_1 G(b_1 x_1) - A_2 F(b_2 x_1) - B_2 G(b_2 x_1) = 0$$
$$B_1 G'(b_1 x_1) - A_2 F'(b_2 x_1) - B_2 G'(b_2 x_1) = 0$$
$$A_2 F(b_2 x_2) + B_2 G(b_2 x_2) - A_3 F(b_3 x_2) - B_3 G(b_3 x_2) = 0$$
$$A_2 F'(b_2 x_2) + B_2 G'(b_2 x_2) - A_3 F'(b_3 x_2) - B_3 G'(b_3 x_2) = 0$$
$$\vdots$$
$$A_{n-1} F(b_{n-1} x_{n-1}) + B_{n-1} G(b_{n-1} x_{n-1}) - B_n G[b_n (x_{n-1} - L)] = 0$$
$$A_{n-1} F'(b_{n-1} x_{n-1}) + B_{n-1} G'(b_{n-1} x_{n-1}) - B_n G'[b_n (x_{n-1} - L)] = 0$$



boundary conditions at $x = x_1$

(1) $\Psi_1(x=x_1) = \Psi_2(x=x_1)$

(2) $\Psi'_1(x=x_1) = \Psi'_2(x=x_1)$

where :

$$\Psi_1(x) = Ae^{ikx} + Be^{-ikx}$$

$$\Psi_2(x) = Ce^{ik'x} + De^{-ik'x}$$

$$k = \sqrt{E - V_0} \quad , \quad k' = \sqrt{V_0 - E}$$

Figure 3. Typical barrier penetration problem

In (19) the primes designate the derivatives of the functions evaluated at the interfaces. System (19) is a homogeneous system involving $2n - 2$ equations and $2n - 2$ unknowns (A_i, $i = 2,3,\ldots,n-1$; B_i, $i = 1,2,3,\ldots,n$). In order to solve for non-trivial solution of the unknowns we must require that the determinant of the coefficients of system (19) be identically zero. That is,

$$|A| = \begin{vmatrix} G(b_1x_1) - F(b_2x_1) - G(b_2x_1) & & & & \\ G'(b_1x_1) - F'(b_2x_1) - G'(b_2x_1) & & & & \\ F(b_2x_2) \quad G(b_2x_2) - F(b_3x_2) - G(b_3x_2) & & & & \\ F'(b_2x_2) \quad G'(b_2x_2) - F'(b_3x_2) - G'(b_3x_2) & & & & \\ \ddots & & & & \\ F(b_{n-1}x_{n-1}) \; G(b_{n-1}x_{n-1}) - G[b_n(x_{n-1} - L)] \\ F'(b_{n-1}x_{n-1}) \; G'(b_{n-1}x_{n-1}) - G'[b_n(x_{n-1} - L)] \end{vmatrix} = 0 \quad . \tag{20}$$

The zeros of this determinant equation are the eigenvalues of the approximate problem, (1), (2) and (3). For each eigenvalue, there is a non-trivial solution for A_i and B_i which in turn defines its corresponding eigenfunction. From this point on, $|A|$ will be looked upon as a function of a real variable E, namely $f(E) = |A|$.

At this point we shall point out differences between system (19) and those obtained by the variational methods, in particular the Rayleigh-Ritz method [1]. In the latter variational method one obtains a homogeneous system in which E is a dependent variable, that is, an algebraic system. If the algebraic system is of order n, one can only obtain n eigenvalues and eigenfunctions. Moreover, there is no definite guarantee that the values of E obtained in this method will be the exact values, as there is no rigorous establishment of convergence in the Rayleigh-Ritz method. On the other hand, the homogeneous system (19) that we derived is a transcendental system. The merits of this system are obvious. The determinant Eq. (20) is a transcendental equation and will always have an infinite number of real roots (zeros). As will be seen later, one can easily define an iterative process whereby we can span a definite energy range, thereby driving the determinant equation to zero at the real eigenvalues encountered in that range. As with any trancendental system, the accuracy desired is only limited by a practical consideration, namely, computational time. The fact that all the roots are real is guaranteed by the fact that (1), (2) and approximation (3) form a Sturm-Liouville system [6]. Equation (1) is a type of Liouville equation whose differential operator

$$H = \frac{d^2}{dx^2} + E - V(x) \tag{21}$$

is a hermitian operator. Indeed, consider two eigenfunctions $u_1(x)$ and $u_2(x)$ being operated upon by H, that is

$$\frac{d^2}{dx^2} u_1 + E_1 u_1 - V u_1 = 0 \quad ,$$

(22)

$$\frac{d^2}{dx^2} u_2 + E_2 u_2 - V u_2 = 0 \quad .$$

Now multiplying on the left the equation for u_1 by u_2 and the equation for u_2 by u_1 and subtracting:

$$(23) \qquad u_2 \frac{d^2}{dx^2} u_1 + u_2(E_1 - V)u_1 - u_1 \frac{d^2}{dx^2} u_2 - u_1(E_2 - V)u_2 = 0 \quad ,$$

or

$$(24) \qquad u_2 \frac{d^2}{dx^2} u_1 - u_1 \frac{d^2}{dx^2} u_2 = u_1 E_2 u_2 - u_2 E_1 u_1 + u_2 V u_1 - u_1 V u_2 = 0 \quad ,$$

but, since

$$(25) \qquad u_2 V u_1 - u_1 V u_1 = u_2 V u_1 - u_2 V u_1 = 0 \quad ,$$

we have

$$(26) \qquad u_2 \frac{d^2}{dx^2} u_1 - u_1 \frac{d^2}{dx^2} u_2 = (E_2 - E_1) u_1 u_2 \quad .$$

The reason we obtain so simple a relation as (26) for comparison is because the modified Liouville Eq. (1) is self-adjoint. This in turn implies that H in (21) is a hermitian operator. It is an easy task to show that the eigenvalues of a hermitian operator are real. Consider the inner product of the two eigenfunctions u and Hu:

$$(27) \qquad (u,Hu) = (u,Eu) = (E^*u,u) = E^*(u,u)$$

where E is the eigenvalue of the hermitian operator H, and E^* is its complex conjugate. Since H is hermitian,

$$(28) \qquad (u,Hu) = (Hu,u) = (Eu,u) = E(u,u) \quad .$$

Subtracting (28) from (27) we obtain

$$(29) \qquad E^*(u,u) - E(u,u) = 0 \quad .$$

Since $(u,u) = 0$, therefore

$$(30) \qquad E^* - E = 0 \quad ,$$

which implies that E is real.

4. THE NUMERICAL METHOD. In this section we shall deal primarily with the algebraic manipulations necessary to implement a numerical solution of Eq. (19). We will first cover the matrix algebra analysis of determinant (20), and then discuss the computation of the eigenfunctions.

A. The Eigenvalue Equation. Equation (20), from now on, will be referred to as the eigenvalue equation. We have already stated that the determinant equation will have an infinite number of real roots that are the bound state eigenvalues for the approximate problem (1), (2) and (3). At this point we will shift the elements in the last column to the second column. The roots of (20) are not affected by this. By doing this, (20) becomes

$$\begin{vmatrix}
G(b_1x_1) & -F(b_2x_1) - G(b_2x_1) & & \\
G'(b_1x_1) & -F'(b_2x_1) - G'(b_2x_1) & & \\
 & F(b_2x_2) \quad G(b_2x_2) & -F(b_3x_2) - G(b_3x_2) & \\
 & F'(b_2x_2) \quad G'(b_2x_2) & -F'(b_3x_2) - G'(b_3x_2) & \\
 & \vdots & & \\
 & -G[b_n(x_{n-1} - L)] & F(b_{n-1}x_{n-1}) \quad G(b_{n-1}x_{n-1}) & \\
 & -G'[b_n(x_{n-1} - L)] & F'(b_{n-1}x_{n-1}) \quad G'(b_{n-1}x_{n-1})
\end{vmatrix} = 0 \quad . \tag{31}$$

We will group the elements in determinant (31) in 2 x 2 sub-matrices as follows:

$$\begin{vmatrix}
\begin{pmatrix} G(b_1x_1) & 0 \\ G'(b_1x_1) & 0 \end{pmatrix} & \begin{pmatrix} -F(b_2x_1)-G(b_2x_1) \\ -F'(b_2x_1)-G'(b_2x_1) \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \cdots & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\
\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} F(b_2x_2) & G(b_2x_2) \\ F'(b_2x_2) & G'(b_2x_2) \end{pmatrix} & \begin{pmatrix} -F(b_3x_2)-G(b_3x_2) \\ -F'(b_3x_2)-G'(b_3x_2) \end{pmatrix} & \cdots & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\
 & \vdots & & & \\
\begin{pmatrix} 0 & -G[b_n(x_{n-1}- L)] \\ 0 & -G'[b_n(x_{n-1}-L)] \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \cdots & \begin{pmatrix} F(b_{n-1}x_{n-1}) & G(b_{n-1}x_{n-1}) \\ F'(b_{n-1}x_{n-1}) & G'(b_{n-1}x_{n-1}) \end{pmatrix}
\end{vmatrix} = 0 \quad . \tag{32}$$

The sub-matrices in the first and last row of first column of determinant (32) correspond to the first and last regions specified by b_1 and b_n and the first and last interfaces specified by x_1 and $x_{n-1} - L$. All other sub-matrices of determinant (32) are of the same form and refer to a single region specified by b_i, $i = 2,3,\ldots,n-1$ and a single interface specified by x_i, $i = 1,2,\ldots,n-1$.

We shall now define the 2 x 2 sub-matrices in determinant (32) as follows:

$$A_{11} = \begin{pmatrix} G(b_1x_1) & 0 \\ G'(b_1x_1) & 0 \end{pmatrix} \qquad A_{n-1,1} = \begin{pmatrix} 0 & -G[b_n(x_{n-1} - L)] \\ 0 & -G'[b_n(x_{n-1} - L)] \end{pmatrix},$$

$$A_{ij} = \begin{pmatrix} F(b_jx_i) & G(b_jx_i) \\ F'(b_jx_i) & G'(b_jx_i) \end{pmatrix}.$$

(33)

A zero will designate the null 2 x 2 sub-matrix. Equation (32) can now be written in a more compact and handable form, namely;

$$|A| = \begin{vmatrix} A_{11} & -A_{12} & 0 & \cdots & & & 0 \\ 0 & A_{22} & -A_{23} & \cdots & & & 0 \\ 0 & 0 & A_{33} & \cdots & & & 0 \\ \vdots & \vdots & \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & A_{n-3,n-3} & -A_{n-3,n-2} & 0 \\ 0 & 0 & 0 & \cdots & 0 & A_{n-2,n-2} & -A_{n-2,n-1} \\ A_{n-1,1} & 0 & 0 & \cdots & 0 & 0 & A_{n-1,n-1} \end{vmatrix}$$

(34) $$= 0 .$$

Post multiplying the last column by $-A^{-1}_{n-1,n-1}A_{n-1,1}$ and adding the result to the first column we obtain

$$
|A| = \begin{vmatrix}
A_{11} & -A_{12} & 0 & \cdots & & 0 \\
0 & A_{22} & -A_{23} & \cdots & & 0 \\
0 & 0 & A_{33} & \cdots & & 0 \\
\vdots & \vdots & \vdots & & & \vdots \\
0 & 0 & 0 & \cdots & A_{n-3,n-3} & -A_{n-3,n-2} & 0 \\
A_{n-2,n-1}A^{-1}_{n-1,n-1}A_{n-1,1} & 0 & 0 & \cdots & 0 & A_{n-2,n-2} & -A_{n-2,n-1} \\
0 & 0 & 0 & \cdots & 0 & 0 & A_{n-1,n-1}
\end{vmatrix}
$$

(35) $= 0$.

Now expanding determinant (35) by the last row

$$
|A| = |A_{n-1,n-1}| \; X
$$

$$
\begin{vmatrix}
A_{11} & -A_{12} & \cdots & & 0 \\
0 & A_{22} & \cdots & & 0 \\
0 & 0 & \cdots & & 0 \\
\vdots & \vdots & & & \vdots \\
0 & 0 & 0 & \cdots & A_{n-4,n-4} & -A_{n-4,n-3} & 0 \\
0 & 0 & 0 & \cdots & 0 & A_{n-3,n-3} & -A_{n-3,n-2} \\
A_{n-2,n-1}A^{-1}_{n-1,n-1}A_{n-1,1} & 0 & 0 & \cdots & 0 & 0 & A_{n-2,n-2}
\end{vmatrix}
$$

(36) $= 0$.

The second determinant on the right-hand side of Eq. (36) has the same form as the original determinant (34). We now repeat the process. Post multiplying the last column of the second determinant on the right-hand side of (36) by $-A^{-1}_{n-2,n-2}A_{n-2,n-1}A^{-1}_{n-1,n-1}A_{n-1,1}$ and adding the result to the first column we obtain

$$|A| = |A_{n-1,n-1}| \times$$

$$\begin{vmatrix} A_{11} & -A_{12} & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 \cdots & -A_{n-4,n-3} & 0 \\ A_{n-3,n-2}A^{-1}_{n-2,n-2}A_{n-2,n-1}A^{-1}_{n-1,n-1}A_{n-1,1} & 0 \cdots & A_{n-3,n-3} & -A_{n-3,n-2} \\ 0 & 0 \cdots & 0 & A_{n-2,n-2} \end{vmatrix}$$

$$= 0 \quad . \tag{37}$$

Again, expanding the second determinant in Eq. (37) by the last row, we obtain

$$|A| = |A_{n-1,n-1}| \times |A_{n-2,n-2}| \times$$

$$\begin{vmatrix} A_{11} & -A_{12} & \cdots & & 0 \\ 0 & A_{22} & \cdots & & 0 \\ 0 & 0 & \cdots & & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 \cdots & A_{n-5,n-5} & -A_{n-5,n-4} & 0 \\ 0 & 0 \cdots & 0 & A_{n-4,n-4} & -A_{n-4,n-3} \\ A_{n-3,n-2}A^{-1}_{n-2,n-2}A_{n-2,n-1}A^{-1}_{n-1,n-1}A_{n-1,1} & 0 \cdots & 0 & 0 & A_{n-3,n-3} \end{vmatrix}$$

$$= 0 \quad . \tag{38}$$

The third determinant on the right-hand side of Eq. (38) has the same form as the original determinant (34). It is evident that we can repeat the process recursively n-2 times to reduce the original eigenvalue Eq. (34) to the form

$$|A| = |A_{n-1,n-1}|X|A_{n-2,n-2}|X|A_{n-3,n-3}|X \cdots X|A_{22}$$

$$|A_{11} + A_{12}A_{22}^{-1}A_{23}A_{33}^{-1}A_{34}A_{44}^{-1} \cdots A_{n-2,n-1}A_{n-1,n-1}^{-1}A_{n-1,1}| = 0 \quad . \tag{39}$$

From Eq. (33) and Table I it is evident that

$$|A_{ii}| = 1 \ , \quad i = 2,3,\ldots,n-1 \quad . \tag{40}$$

Therefore, Eq. (39) reduces to

$$|A| = |A_{11} + A_{12}A_{22}^{-1}A_{23}A_{33}^{-1}A_{34}A_{44}^{-1} \cdots A_{n-2,n-1}A_{n-1,n-1}^{-1}A_{n-1,1}| = 0 \quad . \tag{41}$$

We have, thus, reduced the evaluation of the determinant of a $(2n-2) \times (2n-2)$ matrix in (20) to that of a 2×2 matrix on the right side of (41).

Consider a bound state eigenvalue corresponding to an eigenfunction with two turning points. In Fig. (4) the turning points are indicated at t_1 and t_2. In the classically forbidden regions, the eigenfunction behaves exponentially, that is

$$a_i \equiv E - V_i < 0 \quad . \tag{42}$$

By Table I, the eigenfunctions in these forbidden regions are expressed in terms of hyperbolic functions, reflecting their exponential behaviour. At this point it is necessary to point out that we would run into formidable numerical problems if we were to evaluate Eq. (41) directly using (33). This is because in the classically forbidden region the eigenfunction might be several orders of magnitude smaller than in the allowed region. In numerical methods these problems are known as scaling problems [7], or as asymptotic problems. The reasons can be understood by further examination of Eq. (41). Consider the nonsingular matrix in (41) furthest to the right of the turning point; see also Eq. (33) and Table I.

$$A_{n-2,n-1} = \begin{pmatrix} \cosh b_{n-1}x_{n-2} & \sinh(b_{n-1}x_{n-2})/b_{n-1} \\ b_{n-1}\sinh b_{n-1}x_{n-2} & \cosh b_{n-1}x_{n-2} \end{pmatrix} . \tag{43}$$

In quantum mechanics problems, the arguments of the hyperbolic functions in (43) can grow quite big, say in the neighborhood of 100. We would be in trouble because

$$\sinh 100 = \cosh 100 = e^{100}/2 \quad . \tag{44}$$

When numerically multiplying the 2 x 2 matrices in (41), while evaluating the determinant $A \equiv f(E)$, differences such as sinh 100 - cosh 100 must be evaluated numerically and the computer returns them as zero. That is, the direct evaluation of $A \equiv f(E)$ gives the result

$$A \equiv f(E) \equiv 0 \ , \tag{45}$$

for E in the range of interest. We can remedy this problem by grouping the matrices in (41) except the two singular A_{11} and $A_{n-1,1}$ whose arguments are $b_1 x_1$ and $b_n(x_{n-1} - L)$ respectively (see Eqs. (33)), as follows.

By performing matrix multiplications and using elementary addition formulas for the circular and hyperbolic functions we shall derive expressions for the matrix product $A_{i-1,i}A_{i}^{-1}{}_{,i}$ in the classically forbidden region outside the turning points expressed by

$$a_i \equiv E - V_i < 0 \ , \tag{46}$$

and in the allowed region inside the turning points expressed by (see Fig. 4))

$$a_i \equiv E - V_i > 0 \ . \tag{47}$$

Using (33) we obtain the following matrices:

$$A_{ii} = \begin{pmatrix} F(b_i x_i) & G(b_i x_i) \\ F'(b_i x_i) & G'(b_i x_i) \end{pmatrix} ,$$

$$A_{i-1,i} = \begin{pmatrix} F(b_i x_{i-1}) & G(b_i x_{i-1}) \\ F'(b_i x_{i-1}) & G'(b_i x_{i-1}) \end{pmatrix} . \tag{48}$$

The determinant of A_{ii} is given by

$$|A_{ii}| = 1 \ , \quad i = 2,3,\ldots,n-1 \ . \tag{49}$$

The inverse of A_{ii} is given by

$$A_{ii}^{-1} = \begin{pmatrix} G'(b_i x_i) & -F'(b_i x_i) \\ -G(b_i x_i) & F(b_i x_i) \end{pmatrix}^{T} \frac{1}{|A_{ii}|} \ ,$$

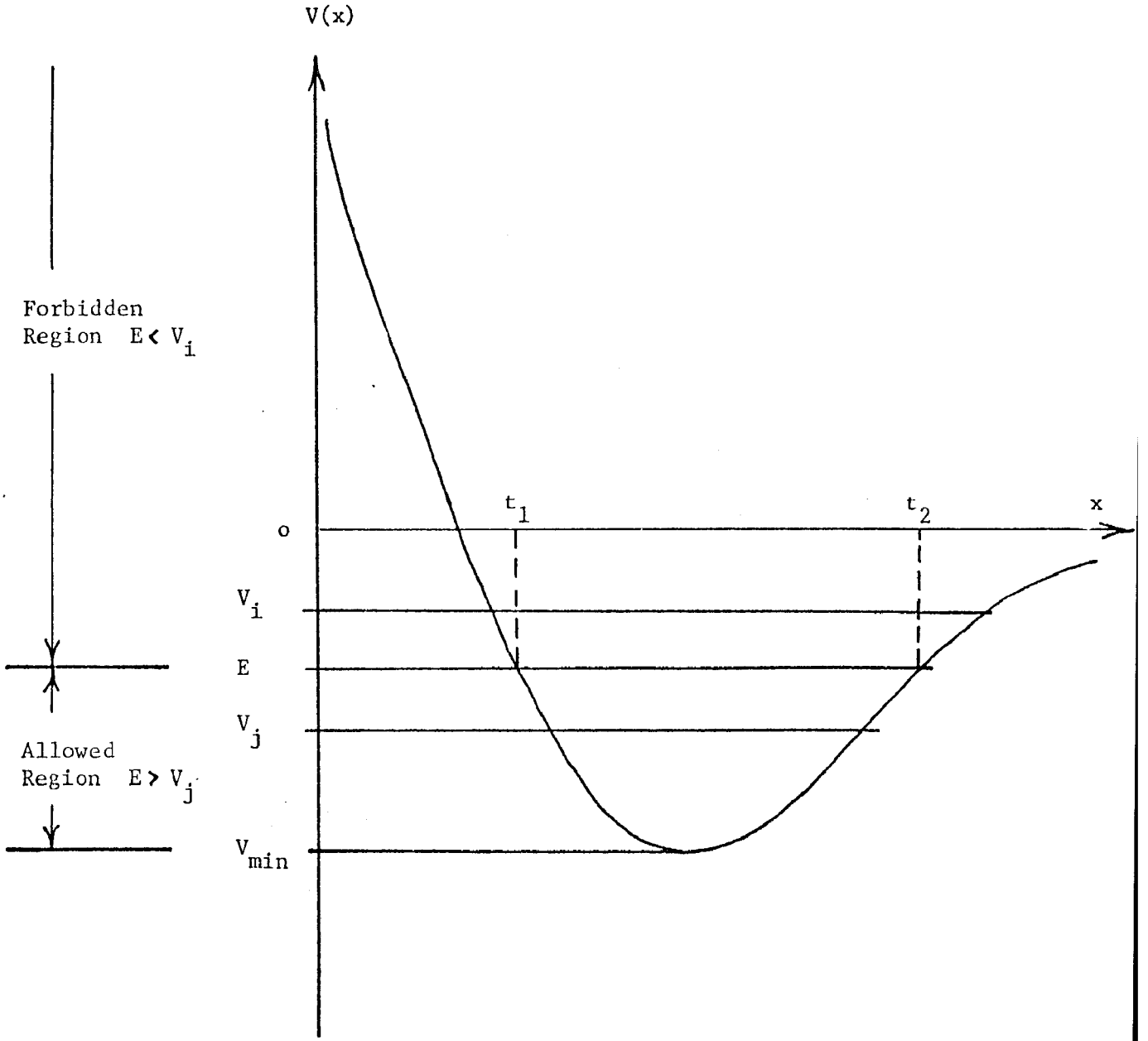


Figure 4. Potential Curve Indicating Turning Points and Their Connection to the Allowed and Forbidden Regions.

$$A_{ii}^{-1} = \begin{pmatrix} G'(b_i x_i) & -F'(b_i x_i) \\ -G(b_i x_i) & F(b_i x_i) \end{pmatrix}^T . \tag{50}$$

The matrix product $A_{i-1,i} A_{ii}^{-1}$ is given by

$$A_{i-1,i} A_{ii}^{-1} = \begin{pmatrix} F(b_i x_{i-1}) & G(b_i x_{i-1}) \\ F'(b_i x_{i-1}) & G'(b_i x_{i-1}) \end{pmatrix} \begin{pmatrix} G'(b_i x_i) & -F'(b_i x_i) \\ -G(b_i x_i) & F(b_i x_i) \end{pmatrix}^T ,$$

$$= \begin{pmatrix} F(b_i x_{i-1})\, G'(b_i x_i) - G(b_i x_{i-1}) F'(b_i x_i) & -F(b_i x_{i-1}) G(b_i x_i) + G(b_i x_{i-1}) F(b_i x_i) \\ F'(b_i x_{i-1}) G'(b_i x_i) - G'(b_i x_{i-1}) F'(b_i x_i) & -F(b_i x_{i-1}) G(b_i x_i) + G'(b_i x_{i-1}) F(b_i x_i) \end{pmatrix} .$$

(51)

For the classically forbidden region outside the turning points (51) becomes

$$A_{i-1,i} A_{ii}^{-1} = \begin{pmatrix} \cosh(b_i (x_i - x_{i-1})) & -b_i^{-1} \sinh(b_i (x_i - x_{i-1})) \\ -b_i \sinh(b_i (x_i - x_{i-1})) & \cosh(b_i (x_i - x_{i-1})) \end{pmatrix} .$$

(52)

And for the allowed region inside the turning points (51) becomes

$$A_{i-1,i} A_{ii}^{-1} = \begin{pmatrix} \cos(b_i (x_i - x_{i-1})) & -b_i^{-1} \sin(b_i (x_i - x_{i-1})) \\ b_i \sin(b_i (x_i - x_{i-1})) & \cos(b_i (x_i - x_{i-1})) \end{pmatrix} . \tag{53}$$

This simple manipulation has, thus, eliminated in one stroke all the scaling difficulties connected with the eigenvalue equation, because the arguments of the hyperbolic functions have been reduced by at least two orders of magnitude with respect to (43). For example, if the potential is approximated by 50 steps, then the step width $x_i - x_{i-1} = h \leq L/50$. The simple, analytical elimination of the scaling difficulties before computation has obvious advantages over a numerical treatment during computation of the eigenvalues.

This section is concluded with a description of the numerical method used to find the roots of the determinant Eq. (41), whose matrices have been grouped as shown in (52) and (53). These roots are approximations to the eigenvalues for the bound states. It will be useful to think of A in (41) as a function of a real variable, f(E),

$$f(E) = |A| \quad , \tag{54}$$

whose zeros will be determined numerically. The eigenvalue search is facilitated by the fact that all eigenvalues are bounded from below by the minimum of the potential [8],

$$E_n \geq V_{min} \quad . \tag{55}$$

In (asymptotic) quantum mechanics problems with deep potential wells, the fundamental energy eigenvalue, E_o, approaches asymptotically the potential minimum [9],

$$E_o \to V_{min} \quad . \tag{56}$$

A computer subroutine EIGEN was written that computes the function f(E) in a predetermined number of integral points in a range

$$V_{min} \leq E \leq V_{right} \quad , \tag{57}$$

where V_{right} is chosen arbitrarily and is to be sufficiently to the right of V_{min} depending on how many eigenvalues are desired. Integral values for V_{min} and V_{right} are chosen. EIGEN computes f(E) for all integral values of E in the range (57). Whenever a change in sign of the function f(E) at two successive (integral) values of E, it will store those two values, and proceeds until it encounters another sign change, whereupon, it will store the two successive (integral) values of E at which the sign change occurs, and so on. This process is continued until the entire range (57) of E is scanned.

Having stored all integral intervals of E for which a sign change for f(E) occurred, the subroutine EIGEN will then subdivide these intervals into ten increments of 0.1 and re-scan each of the intervals for a change in sign of the function f(E) at two successive points. Each one of these integral intervals will contain two successive points for which the function f(E) will change in sign. This latter set of pairs of successive points is stored by EIGEN. Again this last set of intervals is subdivided into ten parts of 0.01 and the process is repeated.

It should now be obvious that this process can be continued indefinitely until the desired accuracy of the eigenvalues is reached. For example, if four iterations are performed, the location of the eigenvalues is ascertained with three decimal accuracy (see the computer output for further examples). The search is terminated when the range (57) is scanned completely or when a predetermined number of eigenvalues has been found.

B. COMPUTATION OF THE EIGENFUNCTIONS. It has been found convenient to rewrite system (19) as

$$
\begin{aligned}
A_{11}\vec{c}_1 - A_{12}\vec{c}_2 &= 0 \\
A_{22}\vec{c}_2 - A_{23}\vec{c}_3 &= 0 \\
A_{33}\vec{c}_3 - A_{34}\vec{c}_4 &= 0 \\
A_{44}\vec{c}_4 - A_{45}\vec{c}_5 &= 0 \\
A_{n-2,n-2}\vec{c}_{n-2} - A_{n-2,n-1}\vec{c}_{n-1} &= 0 \\
A_{n-1,1}\vec{c}_1 + A_{n-1,n-1}\vec{c}_{n-1} &= 0
\end{aligned}
\tag{58}
$$

where A_{ij} are the 2 x 2 matrices given by (33) and $\vec{c}_i$ are the following 2 x 1 vectors:

$$
\vec{c}_1 = \begin{pmatrix} B_1 \\ B_n \end{pmatrix}, \quad \vec{c}_i = \begin{pmatrix} A_i \\ B_i \end{pmatrix}, \quad i = 2,3,\ldots,n-1 \quad . \tag{59}
$$

The components of the vectors $\vec{c}_1$ are the coefficients of the eigenfunctions in the boundary regions 1 and n (see Eq. (13)). The first equation in (58) represents the first two equations in system (19), that is (using Eqs. (33) and (59)):

$$
\begin{pmatrix} G(b_1x_1) & 0 \\ G(b_1x_1) & 0 \end{pmatrix} \begin{pmatrix} B_1 \\ B_n \end{pmatrix} - \begin{pmatrix} F(b_2x_1) & G(b_2x_1) \\ F(b_2x_1) & G(b_2x_1) \end{pmatrix} \begin{pmatrix} A_2 \\ B_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{60}
$$

or, performing the matrix multiplication

$$
\begin{pmatrix} B_1G(b_1x_1) & -A_2F(b_1x_1) - B_2G(b_2x_1) \\ B_1G(b_1x_1) & -A_2F(b_1x_1) - B_2G(b_2x_1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad . \tag{61}
$$

Equation (61) represents the first two equations of system (19) in vector form.

The components of the other vector $\vec{c}_i$, $i \neq 1$, are the coefficients of the eigenfunctions in the inner regions (see Eq. (6)). Once the eigenvalues have been determined, the matrices A_{ij} (Eqs. (33) and (50)) are explicitly determined. A direct numerical solution of (58) would fail because we would encounter scaling problems of a nature mentioned before in the computation of the eigenvalue Eq. (41). These scaling

problems occur while performing numerical solutions of most quantum mechanics problems. In what follows, it shall be demonstrated that a direct numerical solution of (58) would introduce formidable numerical and computational problems.

Let us determine the coefficients of the eigenfunction at the inner region (n - 1) furthest to the right, which is assumed to be outside the turning points t_1 and t_2 in the classically forbidden region (see Fig. 5).

From the last equation in (58), we get

$$A_{n-1,n-1}\vec{c}_{n-1} = -A_{n-1,1}\vec{c}_1 \quad ,$$

or

$$\vec{c}_{n-1} = -A^{-1}_{n-1,n-1}A_{n-1,1}\vec{c}_1 \quad . \tag{62}$$

Let us assume that the eigenfunction is symmetric

$$y(x) = y(-x) \quad , \tag{63}$$

for x in all regions.

Using (63), Eq. (13) and Table I, we have

$$B_1 > 0 \ , \quad B_n < 0 \quad . \tag{64}$$

Since the eigenfunctions in these regions are uniquely determined except by a constant factor, we set

$$B_1 = 1 \ , \quad B_n = -1 \quad . \tag{65}$$

Thus,

$$\vec{c}_1 = \begin{pmatrix} B_1 \\ B_n \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad . \tag{66}$$

Multiplying $A_{n-1,1}$ by $\vec{c}_1$, we get

$$A_{n-1,1}\vec{c}_1 = \begin{pmatrix} 0 & -G[b_n(x_{n-1} - L)] \\ 0 & -G[b_n(x_{n-1} - L)] \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \ ,$$

$$A_{n-1,1}\vec{c}_1 = \begin{pmatrix} G[b_n(x_{n-1} - L)] \\ G[b_n(x_{n-1} - L)] \end{pmatrix} \quad . \tag{67}$$

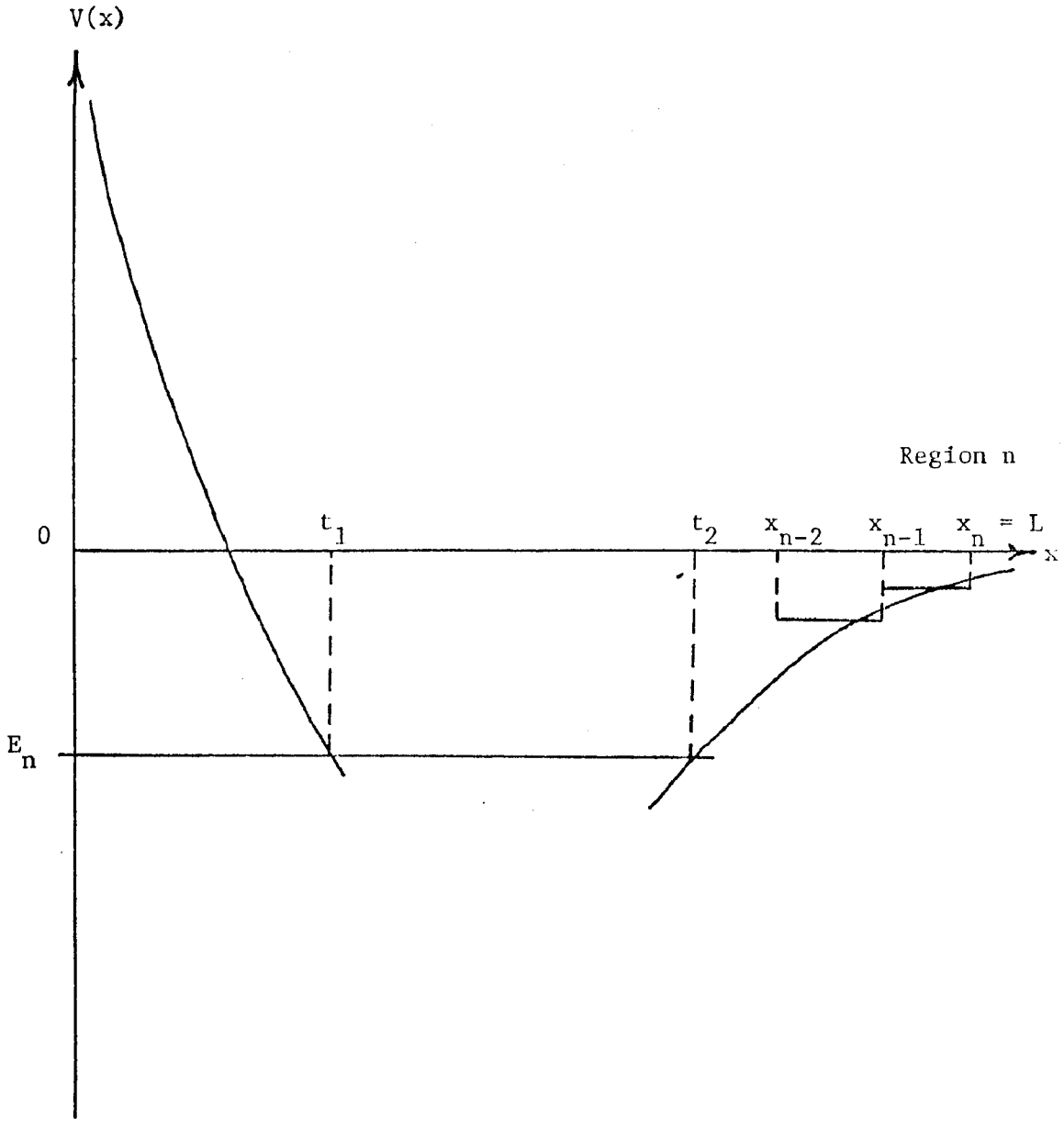


Figure 5. Potential Curve Indicating Inner Regions n - 1, n

Now multiplying $A^{-1}_{n-1,n-1}$ by (67), we get

$$\vec{c}_{n-1} = \begin{pmatrix} \cosh(b_{n-1}x_{n-1}) & b^{-1}_{n-1}\sinh(b_{n-1}x_{n-1}) \\ b_{n-1}\sinh(b_{n-1}x_{n-1}) & \cosh(b_{n-1}x_{n-1}) \end{pmatrix} \begin{pmatrix} G[b_n(x_{n-1} - L)] \\ G[b_n(x_{n-1} - L)] \end{pmatrix},$$

$$\vec{c}_{n-1} = \begin{pmatrix} \cosh(b_{n-1}x_{n-1}) & -b^{-1}_{n-1}\sinh(b_{n-1}x_{n-1}) \\ -b_{n-1}\sinh(b_{n-1}x_{n-1}) & \cosh(b_{n-1}x_{n-1}) \end{pmatrix} \begin{pmatrix} G[b_n(x_{n-1} - L)] \\ G[b_n(x_{n-1} - L)] \end{pmatrix}.$$

(68)

The arguments of the hyperbolic functions in the 2 x 1 vector in (68) are relatively small, while those in the 2 x 2 matrix of (68) are relatively large, typically in excess of 100. There is now no numerical difference between the hyperbolic sine and cosine in the matrix multiplication in (68).

Performing the matrix multiplication in (68) we obtain the coefficients A_{n-1} and B_{n-1} for the eigenfunctions in region n - 1. Using Eq. (6) we obtain the eigenfunction in region n - 1,

$$(69) \quad y(x) = B(\cosh b_{n-1} x - \sinh b_{n-1} x) = 0 , \quad x\varepsilon(x_{n-1}, x_n = L)$$

where B is a constant. Notice that for $b_{n-1} x > 100$ there is no numerical difference between the two hyperbolic functions in (69). Therefore, the computation of the eigenfunction in this region n - 1 would yield zero identically.

This problem can be solved analytically, as was done before with the eigenvalue Eq. (41), by using elementary addition formulas for the circular and hyperbolic functions.

Rewriting the first n - 2 equations of system (58) as:

$$\begin{aligned} \vec{c}_2 &= A^{-1}_{12}A_{11}\vec{c}_1 , \\ \vec{c}_3 &= A^{-1}_{23}A_{22}\vec{c}_2 , \\ \vec{c}_4 &= A^{-1}_{34}A_{33}\vec{c}_3 , \\ &\vdots \\ \vec{c}_{n-1} &= A^{-1}_{n-2,n-1}A_{n-2,n-2}\vec{c}_{n-2} . \end{aligned} \qquad (70)$$

It is obvious that we can now write (integrate) system (70) in terms of the vector c_1 as follows:

$$
\begin{aligned}
\vec{c}_2 &= A_{12}^{-1}A_{11}\vec{c}_1 \ , \\
\vec{c}_3 &= A_{23}^{-1}A_{22}A_{12}^{-1}A_{11}\vec{c}_1 \ , \\
\vec{c}_4 &= A_{34}^{-1}A_{33}A_{23}^{-1}A_{22}A_{12}^{-1}A_{11}\vec{c}_1 \ , \\
&\vdots \\
\vec{c}_m &= A_{m-1,m}^{-1}A_{m-1,m-1}A_{m-2,m-1}^{-1}A_{m-2,m-2}\cdots A_{12}^{-1}A_{11}\vec{c}_1 \ .
\end{aligned}
\tag{71}
$$

The last equation in (71), for $m = n - 1$, determine $\vec{c}_{n-1}$ in terms of $\vec{c}_1$, and the last equation of (58) also express c_{n-1} in terms of c_1. From these two equations we could obtain the last integration constant B_n which appears in $\vec{c}_1$ to completely determine the problem.

For example, from the last equation in (71)

$$
\vec{c}_{n-1} = D\vec{c}_1 \ , \tag{72}
$$

where D is a 2 x 2 matrix designating the product $A_{m-1,m}^{-1}\cdot[A_{m-1,m-1}A_{m-2,m-1}^{-1}\cdots A_{11}]$, and from (56) we get

$$
\vec{c}_{n-1} = E\vec{c}_1 \ , \tag{73}
$$

where E is a 2 x 2 matrix designating the product $-A_{n-1,n-1}^{-1}A_{n-1,1}$. Hence, subtracting the two vector equations we get

$$
\begin{aligned}
\vec{c}_{n-1} - \vec{c}_{n-1} &= (D - E)\vec{c}_1 \\
0 &= \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}\begin{pmatrix} B_1 \\ B_n \end{pmatrix},
\end{aligned}
\tag{74}
$$

or

$$
\begin{aligned}
a_1B_1 + a_2B_n &= 0 \\
a_3B_1 + a_4B_n &= 0
\end{aligned}
\ . \tag{75}
$$

Since B_1 is an arbitrary integration constant, let

$$
B_1 = 1 \ , \quad B_n = -a_1/a_2 \ . \tag{76}
$$

Where a_i, $i = 1,\ldots,4$ designate the matrix elements of the 2 x 2 matrix difference (D - E).

This cannot be readily done because of, again, scaling (asymptotic) problems. That is, numerical errors would accumulate and propagate if (58) were solved from top to bottom; that is the same as starting with an arbitrary value of the eigenfunction at the left boundary and integrating all the way to the right boundary. To avoid this difficulty, the solution of (58) as given by (71) is stopped at m = n/2, i.e., the midpoint of the domain, which is always taken at the minimum (x = a) of the potential function. In symmetric potential problems this is all that is needed, since the eigenfunction is either symmetric or antisymmetric about this point (x = a). For central field problems with asymmetric potentials, after computing the eigenfunction from the left to the midpoint, (58) is solved via (71) from the bottom till the midpoint, i.e., the eigenfunction is now computed from the right inward to the midpoint, and then both pieces of the eigenfunction are matched at the center. The eigenfunctions so computed on either side of the point x = a are the same except that they differ by a constant factor, i.e., their derivatives match at the midpoint.

The matching of the eigenvalues is done by multiplying each piece by its reciprocal value at x = a, so that both pieces also have the same midpoint value, that is,

$$y(x) = \frac{1}{y_L(a)}\, y_L(x) \quad , \quad 0 < x \leq a \quad ,$$

$$y(x) = \frac{1}{y_R(a)}\, y_R(x) \quad , \quad a \leq x < \infty \quad . \tag{77}$$

Except for the first matrix on the right side of (77), the others will be grouped in pairs, as follows:

$$\begin{aligned} A_{ii}A_{i-1,1}^{-1} &= \begin{pmatrix} F(b_ix_i) & G(b_ix_i) \\ F'(b_ix_i) & G'(b_ix_i) \end{pmatrix} \begin{pmatrix} F(b_ix_{i-1}) & G(b_ix_{i-1}) \\ F'(b_ix_{i-1}) & G'(b_ix_{i-1}) \end{pmatrix}^{-1} \quad , \\ &= \begin{pmatrix} F(b_ix_i) & G(b_ix_i) \\ F'(b_ix_i) & G'(b_ix_i) \end{pmatrix} \begin{pmatrix} G'(b_ix_{i-1}) & -F'(b_ix_{i-1}) \\ -G(b_ix_{i-1}) & F(b_ix_{i-1}) \end{pmatrix}^{T} \quad , \\ &= \begin{pmatrix} F(b_ix_i) & G(b_ix_i) \\ F'(b_ix_i) & G'(b_ix_i) \end{pmatrix} \begin{pmatrix} G'(b_ix_{i-1}) & -G(b_ix_{i-1}) \\ -F'(b_ix_{i-1}) & F(b_ix_{i-1}) \end{pmatrix}, \end{aligned} \tag{78}$$

$$= \begin{pmatrix} F(b_ix_i)G'(b_ix_{i-1}) - G(b_ix_i)F'(b_ix_{i-1}) - F(b_ix_i)G(b_ix_{i-1}) + G(b_ix_i)F(b_ix_{i-1}) \\ F'(b_ix_i)G'(b_ix_{i-1}) - G'(b_ix_i)F'(b_ix_{i-1}) - F'(b_ix_i)G(b_ix_{i-1}) + G'(b_ix_i)F(b_ix_{i-1}) \end{pmatrix}.$$

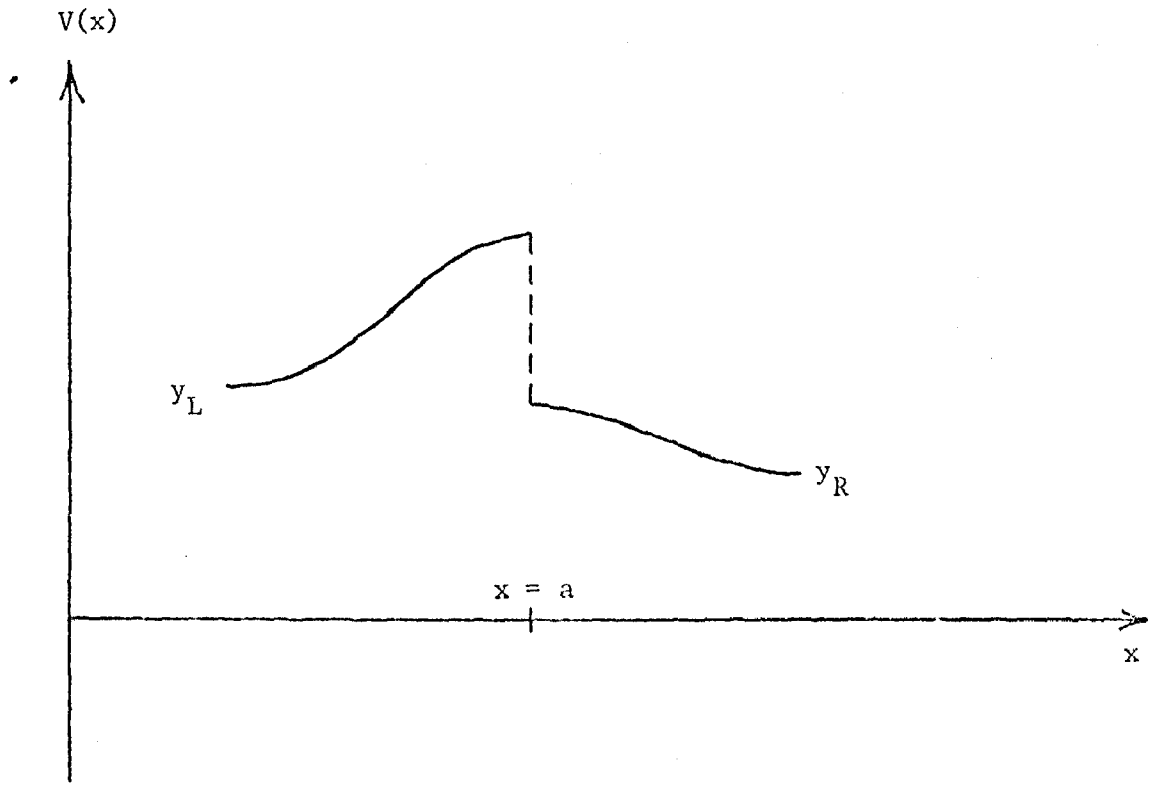


Figure 6. An Eigenfunction Discontinuity at the Midpoint Value of the Domain.

For the classically forbidden region (78) becomes,

$$(79)\qquad A_{ii}A_{i-1,1}^{-1} = \begin{pmatrix} \cosh[b_i(x_i - x_{i-1})] & b_i^{-1}\sinh[b_i(x_i - x_{i-1})] \\ b_i\sinh[b_i(x_i - x_{i-1})] & \cosh[b_i(x_i - x_{i-1})] \end{pmatrix} .$$

For the allowed region (78) becomes

$$(80)\qquad A_{ii}A_{i-1,1}^{-1} = \begin{pmatrix} \cos b_i(x_i - x_{i-1}) & b_i^{-1}\sin b_i(x_i - x_{i-1}) \\ -b_i\sin b_i(x_i - x_{i-1}) & \cos b_i(x_i - x_{i-1}) \end{pmatrix} .$$

Notice that the arguments of the hyperbolic functions in (79) are two orders of magnitude smaller than if the matrices were computed individually. Equations (71) will now be written as follows:

$$(81)\qquad \begin{aligned} \vec{c}_2 &= A_{12}^{-1}\vec{v}_2 , & \vec{v}_2 &= A_{11}\vec{c}_1 , \\ \vec{c}_3 &= A_{23}^{-1}\vec{v}_3 , & \vec{v}_3 &= A_{22}A_{12}^{-1}A_{11}\vec{c}_1 , \\ \vec{c}_4 &= A_{34}^{-1}\vec{v}_4 , & \vec{v}_4 &= A_{33}A_{23}^{-1}A_{22}A_{12}A_{11}\vec{c}_1 , \\ &\vdots & &\vdots \\ \vec{c}_m &= A_{m-1,m}^{-1}\vec{v}_m , & \vec{v}_m &= A_{m-1,m-1}A_{m-2,m-1}^{-1}\cdots A_{11}\vec{c}_1 . \end{aligned}$$

The vectors $\vec{v}_i$ in (81) are evaluated by grouping the matrices in pairs as shown in (79) and (80), except the last matrix A_{11} where the arguments of the hyperbolic functions are $b_i x_i$. As indicated before, the vectors $\vec{v}_i$ involve only one arbitrary integration constant B_1. But we are still in numerical trouble since the inverse matrices A_{ij}^{-1} at the left in (81) still involve large arguments and therefore these inverse matrices are numerically unstable. The computation of the coefficients $\vec{c}_i$ still present numerical difficulties.

We can circumvent this problem since we do not need to compute the vectors of the coefficients $\vec{c}_i$ per se. After the 2 x 1 vectors $\vec{v}_i$ in (81) have been computed they will be of the form

$$(82)\qquad \vec{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix} , \qquad \vec{v}_3 = \begin{pmatrix} v_{31} \\ v_{32} \end{pmatrix} ,$$

$$\vec{v}_m = \begin{pmatrix} v_{m1} \\ v_{m2} \end{pmatrix}.$$

From Eqs. (6), (59) and (81) we have the following expression for the eigenfunctions in vector form or notation:

$$\begin{aligned} y(x) &= \vec{c}_i \cdot \vec{f}(x) \quad , \\ y(x) &= A^{-1}_{i-1,i} \vec{v}_i \cdot \vec{f}(x) \quad , \quad i = 2,3,\ldots,n-1 \end{aligned} \tag{83}$$

where f(x) is the vector function

$$\vec{f}(x) = \begin{pmatrix} f(b_i x) \\ G(b_i x) \end{pmatrix} \quad . \tag{84}$$

Now using Eqs. (13), (82) and (84) we obtain the following expression for the inner product in (83)

$$y(x) = \left[\begin{pmatrix} F(b_i x_{i-1}) & G(b_i x_{i-1}) \\ F'(b_i x_{i-1}) & G'(b_i x_{i-1}) \end{pmatrix}^{-1} \begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} \right]^T \cdot \begin{pmatrix} F(b_i x) \\ G(b_i x) \end{pmatrix} \quad ,$$

$$y(x) = \begin{pmatrix} v_{i1} G'(b_i x_{i-1}) - v_{i2} G(b_i x_{i-1}) \\ -v_{i1} F'(b_i x_{i-1}) + v_{i2} F(b_i x_{i-1}) \end{pmatrix}^T \begin{pmatrix} F(b_i x) \\ G(b_i x) \end{pmatrix},$$

$$y(x) = (v_{i1} G'(b_i x_{i-1}) - v_{i2} G(b_i x_{i-1}), -v_{i1} F'(b_i x_{i-1}) + v_{i2} F(b_i x_{i-1})) \cdot \begin{pmatrix} F(b_i x) \\ G(b_i x) \end{pmatrix}$$

$$\begin{aligned} y(x) = {} & v_{i1} G'(b_i x_{i-1}) F(b_i x) - v_{i2} G(b_i x_{i-1}) F(b_i x) \\ & - v_{i1} F'(b_i x_{i-1}) G(b_i x) + v_{i2} F(b_i x_{i-1}) G(b_i x) \quad , \; i = 2,3,\ldots,n-1 \; . \end{aligned} \tag{85}$$

In the classically forbidden region (85) becomes

$$y(x) = v_{i1}\cosh(b_i x_{i-1})\cosh(b_i x) - v_{i2} b_i^{-1}\sinh(b_i x_{i-1})\cosh(b_i x)$$

$$- v_{i1} b_i \sinh(b_i x_{i-1}) b_i^{-1}\sinh(b_i x) + v_{i2}\cosh(b_i x_{i-1}) b_i^{-1}\sinh(b_i x) ,$$

$$y(x) = v_{i1}\cosh b_i(x_{i-1} - x) - b_i^{-1} v_{i2}\sinh b_i(x_{i-1} - x) ,$$

(86) in region $i = 2,3,\ldots,n-1$.

In the allowed region (85) becomes

$$y(x) = v_{i1}\cos(b_i x_{i-1})\cos(b_i x) - v_{i2} b_i^{-1}\sin(b_i x_{i-1})\cos(b_i x)$$

$$- v_{i1}(-b_i \sin(b_i x_{i-1})) b_i^{-1}\sin(b_i x) + v_{i2}\cos(b_i x_{i-1}) b_i^{-1}\sin(b_i x) ,$$

$$y(x) = v_{i1}\cos b_i(x_{i-1} - x) - b_i^{-1} v_{i2}\sin b_i(x_{i-1} - x) ,$$

(87) in region $i = 2,3,\ldots,n-1$.

In Eqs. (86) and (87) v_{i1} and v_{i2} are components of the vector $\vec{v}_i$ as defined in (81) and (82).

Having been able to use the elementary addition formulas has allowed us to analytically perform the inner product in (83), thus eliminating the need to evaluate hyperbolic functions with very large arguments. It can readily be seen that in the computation of the eigenfunctions in (86) and (87), it is only necessary to evaluate the vectors $\vec{v}_i$ in (81) and not the coefficients $\vec{c}_i$ in (81). The computation of the vectors $\vec{v}_i$ does not present any numerical problems.

5. NUMERICAL RESULTS. In this section we will describe some of the results obtained with the computer subroutine EIGEN. First we dealt with a central field Quantum Mechanics problem, the radial Schrodinger Eq. (1) with Morse's potential [10],

(88) $$V(x) = D(1 - \exp(-a(x - x_e)))^2 - D ,$$

where

(89) $$a = 0.711248 , \quad x_e = 1.9975 , \quad D = 188.4355 .$$

The boundary conditions (2) used are

(90) $$y(0) = y(10) = 0 .$$

The second problem covered was that of Mathieu's Eq. (11), where the fictitious "potential" is

$$V(x) = 2q \cos 2x \quad . \tag{91}$$

The boundary conditions (2) used are in this case

$$y(0) = y(\pi) = 0 \quad . \tag{92}$$

Although Mathieu's equation is not a true Quantum Mechanics problem, Eqs. (1), (91) and (92) can be thought of as the bound states of a particle in a box of length π and infinitely high walls with the potential inside the box given by (91).

The primary reasons why these two problems were chosen are that Morse's potential has well known analytic solutions [12], and provides a good check for the numerical solutions of the eigenvalues; and Mathieu's equation has also been well documented [13] and is a good check for the numerical results obtained for the nodes of the eigenfunctions.

A. Schrodinger's Equation with Morse's Potential. Morse's potential was approximated by a step function with an equal number of steps $m = n/2$ in the ranges

$$0 \leq x \leq 1.9975 \ , \quad 1.9975 \leq x \leq 10 \ . \tag{93}$$

The interface was chosen at the abscissa of the minimum value of the potential

$$V(1.9975) = V_{min} = -188.4355 \quad . \tag{94}$$

This potential varies rapidly in the neighborhood of its minimum than towards the right boundary, where it is relatively flat. The $n/2$ potential steps in the right range of (93) were taken as follows:

$$n/4 \text{ steps in } 1.9975 \leq x \leq 4 \ , \quad n/4 \text{ steps in } 4 \leq x \leq 10 \ .$$

The step function approximation is

$$
V(x) = \begin{cases}
V_1 = (V(0) + V(x_1))/2 \quad , & 0 < x < x_1 \ ; \\
V_2 = (V(x_1) + V(x_2))/2 \quad , & x_1 < x < x_2 \ ; \\
\quad \vdots & \quad \vdots \\
V_m = -188.4355 \quad , & x_{m-1} < x < x_m = 1.9975 \ ; \\
V_{m+1} = -188.4355 \quad , & x_m < x < x_{m+1} \ ; \\
V_{m+2} = (V(x_{m+1}) + V(x_{m+2}))/2 \quad , & x_{m+1} < x < x_{m+2} \ ; \\
\quad \vdots & \quad \vdots \\
V_n = (V(x_{n-1}) + V(x_n))/2 \quad , & x_{n-1} < x < x_n = 10 \ ;
\end{cases}
$$

$$
m = n/2 \ .
$$

(95)

The three step widths used are

$$
0 < x < 1.9975 \ , \quad h_1 = 1.9975/(n/2) \ ,
$$

$$
1.9975 < x < 4 \ , \quad h_2 = (4 - 1.9975)/(n/4) \ ,
$$

$$
4 < x < 10 \ , \quad h_3 = (10 - 4)/(n/4) \ .
$$

The eigenvalues were obtained by searching for the roots of the eigenvalue Eq. (20) in the range

(96) $$-188.0 \leq E \leq -108.0 \ .$$

The numerical results for the first five eigenvalues, when the potential is approximated by n = 200 steps, together with the exact, analytical results are given in Table II. Table III contains the nodes of these first five eigenfunctions.

Table II

First Five Eigenvalues of Schrodinger's Equation
With Morse's Potential

N	E_0	E_1	E_2	E_3	E_4
200	-178.777	-160.264	-142.763	-126.275	-110.796
Exact	-178.799	-160.283	-142.760	-126.288	-110.809

Table III

Nodes of Morse's Eigenfunctions

I	N_{11}	N_{12}	N_{13}	N_{14}	N_{15}
1	0.000	0.000	0.000	0.000	0.000
2	1.888	0.000	0.000	0.000	0.000
3	1.768	0.000	0.000	0.000	0.000
4	1.688	0.000	0.000	0.000	0.000

N_{ij} means the J_{th} node of the I_{th} eigenfunction.

B. Mathieu's Equation. The Mathieu's "potential" is approximated as follows:

$$V(x)=\begin{cases} V_1 = V(0) = 2q \quad , & 0 < x < x_1 \quad ; \\ V_2 = (V(x_1) + V(x_2))/2 \quad , & x_1 < x < x_2 \quad ; \\ \vdots & \vdots \\ V_m = -2q \quad , & x_{m-1} < x < x_m = \pi/2 \quad ; \\ V_{m+1} = -2q \quad , & x_m < x < x_{m+1} \quad ; \\ \vdots & \vdots \\ V_n = V(x_n) = 2q \quad , & x_{n-1} < x < x_n = \quad ; \end{cases}$$

$$m = n/2 \quad .$$

(97)

The potential has the minimum at the center

$$V_{min} = V(\pi/2) = -2q \quad . \tag{98}$$

The eigenvalues were obtained by searching for the roots of the eigenvalue Eq. (20) in the range

$$-70.0 \leq E \leq 30.0 \quad . \tag{99}$$

The numerical results for q = 40 are given in Tables IV and V, for n = 200, together with the exact, analytical results taken from Ince's paper [13].

Table IV

First Five Eigenvalues of Schrodinger's Equation
With Mathieus Potential

N	E_0	E_1	E_2	E_3	E_4
200	-67.595	-43.342	-20.200	1.736	22.337
Exact	-67.606	-43.352	-20.208	1.730	22.332

Table V

Nodes of Mathieu's Eigenfunctions

I	N_{11}	N_{12}	N_{13}	N_{14}	N_{15}
1	1.563	2.757	3.118	0.000	0.000
2	1.359	1.767	2.882	0.000	0.000
3	1.202	1.563	1.924	3.071	3.118
4	1.060	1.406	1.720	2.066	3.087

N_{ij} means the J_{th} node of the I_{th} eigenfunction.

REFERENCES

1. P. M. Morse and H. Feshbach, "Method of Theoretical Physics," Vol. II, p. 1117, McGraw-Hill Book Company, Inc., New York, 1953.

2. J. W. Cooley, Math. Comp. 15 (1961), 363.

3. B. Wendroff, "Theoretical Numerical Analysis," p. 23, Academic Press, New York, 1966.

4. Reference 1, p. 1092.

5. L. Schiff, "Quantum Mechanics," 2nd edition, p. 184, McGraw-Hill Book Company, Inc., New York, 1955.

6. Reference 1, Vol. I, p. 719.

7. J. M. Blatt, J. Comp. Phys. 1 (1967), 382

8. R. Courant and J. Hilbert, "Methods of Mathematical Physics," Vol. I, p. 445, Wiley (Interscience), New York, 1962.

9. J. Canosa and J. D. Cole, J. Math. Phys. 9 (1968), 1915.

10. Reference 5, p. 304.

11. N. W. McLahlan, "Theory and Application of Mathieu Functions," Dover Publications, Inc., New York, 1964.

12. P. M. Morse, Phys. Rev. 34 (1929), 57.

13. E. L. Ince, Proc. Roy. Soc. Edinburgh Sect. A. 52 (1931-32), 355. Also ibid., 52 (1931-32), 424.

LIST OF ATTENDEES

21st Conference of Army Mathematicians

14-16 May 1975

NAME	ADDRESS
BAYLOT, E. A.	US Army Engineers Waterways Experiment Station Box 631, Vicksburg, Mississippi 39180
BOWDEN, Charles M.	Quantum Physics, Physical Sci Directorate, US Army Missile Command, Redstone Arsenal, AL 35809
BRANDSTEIN, Alfred G. DR.	Harry Diamond Laboratories, ATTN: AMXDO-EM 2800 Powder Mill Rd, Adelphi, MD 20783
BRATT, Howard M.	Eustis Directorate, US Army Air Mobility R&D Lab, ATTN: SAVDL-EU-MOR, Ft. Eustis, VA 23604
CARASSO, Alfred	Math Research Center, Univ of Wisconsin, Madison, Wisconsin, 53706
CARROLL, Robert M. DR.	Army Research Institute, Arlington, Virginia
CHANDRA, Jagdish DR.	U. S. Army Research Office, Box CM, Duke Station, Durham, N.C. 27703
CHU, S. C.	General Thomas J. Rodman Laboratories Rock Island Arsenal, Illinois
CHUI, C. K.	Dept of Mathematics, Texas A & M University, College Station, Texas 77843
CLIBORN, Robert I.	P.O. BOX 6057, Fort Bliss, Texas 79916
CLODFELTER, Michael J.	US Army Harry Diamond Laboratories, 2800 Powder Mill Road, Adelphi, MD 20783
COHEN, Donald S. PROF	Applied Mathematics 101-50, California Institute of Technology, Pasadena, California 91109
DAVIS, Paul Wm. DR.	US Army Research Office, Box CM Duke Station, Durham, NC 27706
ELDER, Alexander S.	Interior Ballistics Lab, US Army Ballistic Research Laboratories, Aberdeen Proving Ground, MD 21005

NAME	ADDRESS
CULPEPPER, Gideon	Army Missile Test & Evaluation Directorate
DE KINDER, R.E.	US Army Tradoc Systems Anal Activity
DYLLA, F. T.	Army Missile Test & Evaluation Directorate
ENGEBOS, Bernard	Atmospheric Science Laboratories
ENGELHARDT, H. W.	US Army Tradoc Systems Anal Activity
FEENEY, James K.	Army Missile Test & Evaluation Directorate
FIELD, Edward	Instrumentation Directorate
FIELDER, Gladys S.	National Range Operations Directorate
FIERRO, Roberto	National Range Operations Directorate
FUENTES, F.	US Army Tradoc Systems Anal Activity
GADNEY, George	National Range Operations Directorate
GIBSON, Jon	National Range Operations Directorate
GRAVES, J. A. DR.	US Army Tradoc Systems Anal Activity
GREER, William J.	Army Missile Test & Evaluation Directorate
GROVER, Ken	National Range Operation Directorate
GUTIERREZ, Florentino	US Army Tradoc Systems Anal Activity
HAFEN, J. A.	US Army Tradoc Systems Anal Activity
HEATH, Roy	National Range Operations Directorate
HIGGINS, Patrick J.	National Range Operations Dir, Anal & Cmpt Div
HORTON, Terry W.	National Range Operations Directorate
HOWELL, R. W. 1LT	US Army Tradoc Systems Anal Activity
JENKINS, Jefferson	National Range Operations Directorate
KING, Roger	Army Missile Test & Evaluation Directorate
KNAPP, Donald	National Range Operations Directorate

NAME	ADDRESS
GREVILLE, Thomas N. E.	Mathematics Research Center, Univ of Wisconsin 2022 Baltimore Rd, Apt L-23, Rockville, MD 20851
HARRIS, Paul DR.	Concepts & Effectiveness Div, NDED Picatinny Arsenal, New Jersey 07801
HERSHNER, Ivan R. JR.	Room 3E-365, The Pentagon, ATTN: HQDA, DCSRDA Washington, D.C. 20310
HOFFMAN, Alan J.	IBM Research Center, Mathematical Sciences, P.O. Box 218, Yorktown Heights, New York 10598
HUANG, Nai-Chien	Mathematics Research Center, Univ of Wisconsin 610 Walnut Street, Madison, Wisconsin 53706
HUANG, Y. K.	Watervliet Arsenal, Watervliet, New York 12189
HUSSAIN, Moayyed A.	Benet Weapons Laboratories, Watervliet Arsenal, Watervliet, New York 12189
HWANG, John D.	HQ, US Army Air Mobility R&D Laboratory, Ames Research Center, Moffett Field, CA 940 35
JORGENSEN, Charles DR.	HUMRRO/ARI, Bldg 118, Ft Bliss, Texas 79916
KURKJIAN, Badrig M. DR.	US Army Materiel Command, ATTN: AMCRD-R 5001 Eisenhower Ave, Alexandria, VA 22333
LAZARUK, John L. DR.	Comptroller, S&EA Division, HQ, USACC, Fort Huachuca, Arizona 85613
LEESE, Eric	Dept of National Defense, Operational Research Anal Establishment, Ottawa, Ontario K1H 6H8
LEHNIGK, Siegfried H.	Physical Sciences Dir, US Army Missile RD&E Lab, US Army Missile Command, Redstone Arsenal, AL 35809
LEJEUNE, David W. COL.	US Army Communications Command Fort Huachuca, Arizona 85613
McNAULTY, Mr.	Math Research Center, Univ of Wisconsin Madison, Wisconsin 53706
McINTRYRE, Robert DR.	University of Texas at El Paso, El Paso, TX
MESSINGER, Martin DR.	Effectiveness Branch, C&ED, ADED, Picatinny Arsenal, Dover, NJ 07081

NAME	ADDRESS
LAMB, Judith	National Range Operations Directorate
LEATH, Robert	Army Missile Test & Evaluation Directorate
McLAUGHLIN, Dale	National Range Operations Directorate
MEYER, Jerry	National Range Operations Directorate
MORALES, George	Army Missile Test & Evaluation Directorate
NEWTON, G. H.	US Army Tradoc Systems Anal Activity
OGAZ, Juan	National Range Operations Directorate
PAPPAS, James	National Range Operations Directorate
POTTER, G. A.	US Army Tradoc Systems Anal Activity
REDE, E. JR.	US Army Tradoc Systems Anal Activity
REXRODE, Doyle D.	Army Missile Test & Evaluation Directorate
RIPPY, F. E.	US Army Tradoc Systems Anal Activity
RODRIGUEZ, Pedro	National Range Operations Directorate
SANCHEZ, Ernest J.	National Range Operations Directorate
SASSEMFELD, Helmut DR.	US Army Tradoc Systems Anal Activity
SCHRAMM, R. C.	US Army Tradoc Systems Anal Activity
SHEPHERD, William L.	Instrumentation Directorate
SOUTHWORTH, E. F.	Army Missile Test & Evaluation Directorate
TURNER, R. H.	National Range Operations Directorate
VALENCIA, Robert	National Range Operations Directorate
WOOD, R. L. MAJ	US Army Tradoc Systems Anal Activity
ZUNIGA, G.	US Army Tradoc Systems Anal Activity

NAME	ADDRESS
McJUNKIN, Harry	El Paso, Texas
MITTENTHAL, Lothrop COL	US Army Research Office, P.O. BOX 12211 Research Triangle Park, North Carolina 27709
MORRISON, Clyde A. DR.	Harry Diamond Laboratories, Branch 320, 2800 Powder Mill Road, Adelphi, MD 20783
MURRILLO, Joe R.	ADB, Fort Bliss, Texas
NOBLE, Ben PROF	Mathematics Research Center, Univ of Wis, 610 Walnut Street, Madison, Wisconsin 53706
PELL, William H. DR.	National Science Foundation, Math Sciences Sec, 1800 G. St, N. W, Rm 305, Washington, D.C. 20550
POLK, J. F.	Applied Mathematics & Sciences Lab, Ballistic Research Lab, Aberdeen Proving Ground, MD 21005
POLLIN, Jack M. COL.	Dept of Mathematics, US Military Academy West Point, New York 10996
PROVENCIO, Jesus PROF	University of Texas at El Paso, El Paso, TX
PURYEAR, Van A.	Troop Support Command, 4300 Goodfellow, St. Louis, Missouri 63120
RALL, Louis B. PROF	Mathematic Research Center, Univ of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706
RAMIREZ, Enrique LTC	AFOSR/NM, 1400 Wilson Blvd, Arlington, VA 22209
ROSS, Edward W. JR. DR.	US Army Natick Development Center, Kansas Street, Natick, MA 01760
ROSSER, J. Barkley	Mathematic Research Center, Univ of Wisconsin 610 Walnut Street, Madison, Wisconsin 53706
SINGLETON, Robert E.	US Army Research Office, Engineering Sciences Div Box CM, Duke Station, Durham, NC 27706
SMITH, P. W.	Dept of Mathematics, Texas A & M University College Station, Texas 77843
STEEVES, Earl C.	US Army Natick Development Center Natick, Massachusetts 01760
TAGAKI, Shunsuke DR.	US Army Cold Regions Research & Engineering Lab, P.O. Box 282, Hanover, New Hampshire 03755
THOMPSON, James L.	US Army Tank-Automotive Command, ATTN: AMSTA-RHMM Warren, Michigan 48090

NAME	ADDRESS
TOAL, J. J.	General Thomas J. Rodman Laboratory, Rock Island Arsenal, Rock Island, Illinois 61201
ULLRICH, George W.	USAMERDC, Countermine/Counter Intrusion Dept, Special Projects Div, Ft. Belvoir, VA 22060
VITERBI, Andrew J. DR.	LINKABIT Corporation, 10453 Roselle Street San Diego, California 92121
WEISS, Richard A.	Soil and Pavements Laboratory, US Army Engineers Waterways Experiment Station, Vicksburg, MS 39180
WILBURN, John B. Jr.	U. S. Army Electronic Proving Ground, Ft. Huachuca, Arizona 85613
WILLIS, Roger F.	US Army Combined Arms Combat Development Activity, Ft. Leavenworth, Kansas 66048
WU, Julian J. DR.	Benet Laboratories, Watervliet Arsenal, Watervliet, New York 12189
YALAMANCHILI, Rao V. S.	General Thomas J. Rodman Laboratory, Rock Island Arsenal, Rock Island, Illinois

WHITE SANDS MISSILE RANGE ATTENDEES

AGEE, W. S.	National Range Directorate
ALVAREZ, L.	US Army Tradoc Systems Anal Activity
BAAS, Elwood D.	Army Missile Test & Evaluation Directorate
BAKER, Charles	National Range Operations Directorate
BECHTLOFFT, D. L.	US Army Tradoc Systems Anal Activity
BRANTLEY, L. W.	US Army Tradoc Systems Anal Activity
BUTLER, W. G.	US Army Tradoc Systems Anal Activity
CARIOLA, Eugene	US Army Tradoc Systems Anal Activity
CARRILLO, J. E.	US Army Tradoc Systems Anal Activity
CARTNER, D. C.	US Army Tradoc Systems Anal Activity
CASTILLO, Cesar	National Range Operations Directorate
CASTRO, Oscar J.	Army Missile Test & Evaluation Directorate
CHAVEZ, Chris	National Range Operations Directorate

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ARO Report 76-1	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: TRANSACTIONS OF THE TWENTY-FIRST CONFERENCE ON ARMY MATHEMATICIANS	5. TYPE OF REPORT & PERIOD COVERED: Interim Technical Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Army Mathematics Steering Committee on Behalf of the Chief of Research, Development and Acquisition	12. REPORT DATE: February 1976
	13. NUMBER OF PAGES: 730
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: U. S. Army Research Office AMXRO P. O. Box 12211 Research Triangle Park, NC 27709	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited. The findings in this report are not to be considered as an official Department of the Army position; unless so designated by other authorized documents.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

This is a technical report resulting from the Twenty-first Conference of Army Mathematicians. It contains most of the papers on the Agenda of this meeting. These treat various Army applied mathematical problems.

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

graph theory
stability
buckling
analytic functions
dynamic models
hybrid analog-simulation computer models
infinitesimal transformation groups
polynomial approximation
Gronwall-Reid-Bellman lemma
Volterra integral operators
nonlinear problems
elastic yarns
digital communications
shear forces
difference solutions
electromagnetic surface waves
opto-electronic materials
collision coordinates
ballistic environment
shock theory
subsonic flow
backward beam equation
parabolic equation
in-flight reliability
trajectory estimation
proving programs correct
plastic flow
diffusive systems

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified